Contract No:

This document was prepared in conjunction with work accomplished under Contract No. DE-AC09-08SR22470 with the U.S. Department of Energy (DOE) Office of Environmental Management (EM).

Disclaimer:

Machine Learning Using Open Data Sources for Detection of Nuclear Proliferation Activities (U)

Savannah River National Laboratory:

T. L. Danielson
J. A. Pike
T. Whiteside

Discovery Analytics Center at Virginia Polytechnic Institute and State University:

B. Mayer
N. Muralidhar
N. Self
P. Butler

January 2021

SRNL-STI-2021-00047, Revision 0

DISCLAIMER

Printed in the United States of America

Prepared for
U.S. Department of Energy

SRNL-STI-2021-00047
Revision 0

Keywords: *Artificial Intelligence*
Machine Learning
Proliferation

Retention: *Permanent*

Machine Learning Using Open Data Sources for Detection of Nuclear Proliferation Activities (U)

Savannah River National Laboratory:

T. L. Danielson
J. A. Pike
T. Whiteside

Discovery Analytics Center at Virginia Polytechnic Institute and State University:

B. Mayer
N. Muralidhar
N. Self
P. Butler

January 2021

Prepared for the U.S. Department of Energy under contract number DE-AC09-08SR22470.

Savannah River National Laboratory®
OPERATED BY SAVANNAH RIVER NUCLEAR SOLUTIONS

REVIEWS AND APPROVALS

AUTHORS:

T. L. Danielson, Environmental Sciences and Dosimetry Date

J. A. Pike, Advanced Modelling, Simulation and Analytics Date

T. Whiteside, Nuclear Measurements Date

B. Mayer, Discovery Analytics Center at Virginia Polytechnic Institute and State University Date

N. Muralidhar, Discovery Analytics Center at Virginia Polytechnic Institute and State University Date

N. Self, Discovery Analytics Center at Virginia Polytechnic Institute and State University Date

P. Butler, Discovery Analytics Center at Virginia Polytechnic Institute and State University Date

APPROVALS:

D. G. Jackson, Jr., Manager, Environmental Sciences and Dosimetry Date

C. M. Gregory, Manager, Nuclear Measurements Date

P. L. Lee, Manager, Advanced Modelling, Simulation and Analytics Date

EXECUTIVE SUMMARY

In FY2020, Savannah River National Laboratory (SRNL) in collaboration with the Sanghani Center for Artificial Intelligence and Data Analytics (SCAIDA) at Virginia Polytechnic Institute and State University (VT) and funded by the Department of Energy's (DOE) Defense Nuclear Nonproliferation Research and Development, began developing a demonstration prototype system that uses multiple machine learning and data analytic methods on large-scale open data sources to identify new, developing, and/or undeclared nuclear programs. Using the announcement in May 2018 of the proposed Savannah River Plutonium Processing Facility (SRPPF) as a test subject, the goal of this 2-year project is to forecast the SRPPF using only data prior to May 2018. The project work is split into a preliminary prototype development for the first year with an initial evaluation of viability followed by the second year of development to create an integrated prototype system and more extensive performance evaluation. This report documents the results of the preliminary-phase tasks.

Two large data archives were used to create test datasets using a complex query method with a semi-supervised generation of query terms. This method represents the initial signal concentration step and eliminates a great deal of noise without compromising the sparse and widely distributed signal in the data. One data source representing social media includes nearly 4 years of Twitter™[a] data. The other data source, Webhose Ltd. (https://webhose.io/), provides historical archive to a broad spectrum of media postings (articles, news, blogs, etc.).

Models and methods were developed for each data archive to detect organization names, facility, and function identification as well as events within the openly available data.

Analysis of Twitter™ data demonstrates that word embedding models trained on data obtained from queries using the glossary of terms produce an interpretable signal of key terms and the connectedness to relevant entities that automated models can use. The research team is still evaluating current methods to determine the best method for use in the prototype system. For example, a time dependent word embedding algorithm that retains all cumulative data may be advantageous in cases that create a strong link to an event that occurs earlier or later in time. However, keeping all data could create too low of a "signal-to-noise" ratio and prohibit straightforward detection of changes of interest. A "rolling window" can help to eliminate noisy data by shifting over time, however, the window size or rolling length needs tuning to avoid missing key events that are linked across time. The research team will use this analysis to build a prototype system that can identify the indicators of nuclear activity.

For the news article data, elements of an automated preliminary data pipeline have been built and the evaluation of each step showed strong relationships and indication of entity detection while an anomaly detection model that analyzes the temporal evolution of networks is still under development.

Each algorithm has presumed advantages and disadvantages, as well as parameters that need to be tuned and characterized for application in the prototype model development. The second year of the project development will develop fusion models to leverage the advantages of the individual models. A prototype system will be created and demonstrated in the next year of the project.

[a] Twitter is a trademark of Twitter, Inc. or its affiliates.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DOD	Department of Defense
DOE	Department of Energy
EMBERS	Early Model Based Event Recognition using Surrogates
MOX	Mixed Oxide Fuel
NER	Named Entity Recognition
NN	Nearest Neighbor
NNSA	National Nuclear Security Agency
NRC	Nuclear Regulatory Commission
NTI	Nuclear Threat Initiative
PCA	Principal Component Analysis
SCAIDA	Sanghani Center for Artificial Intelligence and Data Analytics
SME	Subject Matter Expert
SNL	Sandia National Laboratory
SRNL	Savannah River National Laboratory
SRPPF	Savannah River Plutonium Processing Facility
SRS	Savannah River Site
US	United States
VT	Virginia Polytechnic Institute and State University

1.0 Introduction

In FY2020 the Savannah River National Laboratory (SRNL) in collaboration with the Sanghani Center for Artificial Intelligence and Data Analytics (SCAIDA) at Virginia Polytechnic Institute and State University (VT), and funded by the Department of Energy's (DOE) Defense Nuclear Nonproliferation Research and Development, began developing a demonstration prototype system that uses multiple machine learning and data analytic methods on large-scale open data sources to identify new, developing, and/or undeclared nuclear programs. The 2-year project was split into a preliminary prototype development for the first year with an initial evaluation of viability followed by the second year of development to create an integrated prototype system and more extensive performance evaluation. This report documents the results of the preliminary-phase tasks and viability evaluation.

1.1 Background

The objective of the project is to apply the well-developed data analytics technologies and know-how of VT's Sanghani Center with nonproliferation application guidance from SRNL and other experts to demonstrate the viability of using multiple data analytics methods on large-scale open data sources to detect new, developing, or undeclared nuclear programs. The prototype development is focused on demonstrating the ability to identify a new program/facility before known publicly, specifically, the proposed Savannah River Plutonium Processing Facility (SRPPF) which is also commonly referred to as the pit production facility. Generically, the facility fits the weapons development process as fissile core fabrication. DOE's National Nuclear Security Agency (NNSA) announced this project on May 10, 2018 via a release "Joint Statement from Ellen M. Lord and Lisa E. Gordon-Hagerty on Recapitalization of Plutonium Pit Production"

> *To achieve DoD's 80 pits per year requirement by 2030, NNSA's recommended alternative repurposes the Mixed Oxide Fuel Fabrication Facility at the Savannah River Site in South Carolina to produce plutonium pits while also maximizing pit production activities at Los Alamos National Laboratory in New Mexico. (Joint Statement from Ellen M. Lord and Lisa E. Gordon-Hagerty on Recapitalization of Plutonium Pit Production 2018)*

The facility is planned to start up in 2030 which offers two possible courses of development, one based on open source data and predictability before the announcement and the other based on accumulation of data after the announcement that would detail the many related events that could be used for developing detailed related event definitions for a forecasting system. In this project we will focus first on predictability before the announcement and consider how to accumulate information for the latter as time and resources allow.

In addition, a semi-blind test of the methods and algorithms are planned by testing translatability to additional parts of the weapon development process. The methods developed in this prototype system are planned to be extended to fuel reprocess for finding indications of new reprocessing events and specifically for initiation of a new reprocessing process at SRS, a site already known for large scale fuel reprocessing. Such a test may allow characterization of relative sensitivity of the methods and models.

1.2 Known State of the Art

In the broad context of data science, the identification and forecasting of previously unknown proliferation activities falls under a complementary grouping of sub-categories, namely: anomaly detection, change detection, and event detection. Several well-developed theoretical foundations (e.g., statistical and graph theoretical) and techniques (e.g., nearest-neighbor, classification, and clustering) exist for each of these sub-categories. Ranshous et al. provides a survey of anomaly detection in dynamic networks with an emphasis on graph theoretical formalisms that are widely applicable to the proposed work since a diverse

grouping of open data sources will be used (Ranshous, et al. 2015). Application of such approaches are shown in work related to identifying events like breaking news (Hu, et al. 2012), mass protests (Zhao, et al. 2014), and disease outbreaks (Rekatsinas, et al. 2015). Social media has played a key role in many of these network-based event detection methodologies. Ahmad et al. (Ahmad, et al. 2017) provides a summary of a more statistically based unsupervised anomaly detection approach that processes data in real-time to detect anomalies within interconnected data streams and provides benchmarking for several related algorithms. Statistical approaches to identifying anomalous patterns have also been used to predict rare weather events and earthquakes (Olson, et al. 2011) and change detection (Lunetta, et al. 2006) approaches have been used to identify changes in vegetation index from normalized difference vegetation index to identify different land-use patterns. Another critical part of anomaly and event detection is the ranking and filtration pipeline which allows the focus to be shifted to highly relevant articles for a particular task by eliminating irrelevant noise in the data (Croft, Metzler and Strohman 2010).

An extensive body of literature is available in the context of mining and processing open data sources, especially news articles, blogs, and social media. Key techniques used in this work are: dynamic query expansion algorithms to fetch relevant data in an unsupervised fashion; association analysis and graph theoretical concepts for forming connections between people and popular/recurring thoughts and ideas; linguistic processing algorithms that overcome the subtleties of natural language in order to evaluate the relevance and/or understand sentiments; and traditional logistic regression machine learning, transfer learning, Bayesian inference, and maximum likelihood estimate models that establish and confirm connections and make structured predictions regarding a particular event or collection of events.

Since the start of the project, researchers have become aware of other work that is topically aligned with this project's objectives, most notably, work by the Nuclear Threat Initiative (NTI) and the Center for Advanced Defense Studies (Arterburn, Dumbacher and Stoutland 2021). NTI reported in January 2021 on a pilot project to demonstrate the viability of using open source data and machine learning to detect high-risk and/or illicit nuclear trade. Though similar in the general ideas, the technical application and approach are different. The NTI project focused on indicators of illicit trade which is one aspect of consideration that would be complimentary to the objective of this project. Furthermore, the NTI reported success in observing indicators in the data suggests that the indicators observed in this work are likely to become more robust as the second half of this project is worked with reasonable probability of translatability to other parts of the world.

Precedent for the general type of forecasting technology that is of interest has been established through full deployment of programs such as: Integrated Crisis Early Warning System (O'Brien 2010), originally funded by Defense Advanced Research Project Agency and developed and maintained by Lockheed Martin; Political Instability Task Force (Goldstone, et al. 2010), funded by the US Central Intelligence Agency; and Early Model Based Event Recognition using Surrogates (EMBERS) (Ramakrishnan, et al. 2014) developed within the Sanghani Center (https://sanghani.cs.vt.edu/) and funded by Intelligence Advanced Research Project Agency. The common thread between these programs is that they are developed to use large data sets (often open source data sets) to forecast civil unrest and support decisions regarding the potential allocation of resources to mitigate and/or respond to crises. While the respective implementations of each of these programs have had measurable successes, the EMBERS system is at the forefront of the state-of-the-art in terms of its spatio-temporal resolution and the accuracy and specificity of the events predicted.

This project addresses a key technical gap by demonstrating the feasibility of using such techniques for making predictions regarding obscure events that unravel over longer time periods of several months to years.

1.3 Approach

The prototype system under development adapts the modular architecture of the EMBERS forecasting system (Ramakrishnan, et al. 2014) that is shown in Figure 1-1 with models developed specifically for detection of indicators for the long evolution of weapons development facilities and programs.

Data sources could include multiple media types, though this initial development focuses on text or readily convertible data to text to facilitate speedy initial system development. Details of how each data source are ingested and enriched along with predictive model development are described in their respective sections. Each section describes initial methods and models that have been developed and tested by data source. Additional sources and data types may be added in subsequent development.

The second half of this project will assemble the individual parts into the integrated structure similar to that of Figure 1-1. Modifications will be made to the architecture as needed to ensure the process is most effective for the task at hand and efficiently uses the models. At the time of this report, fusion and prediction models are under development.

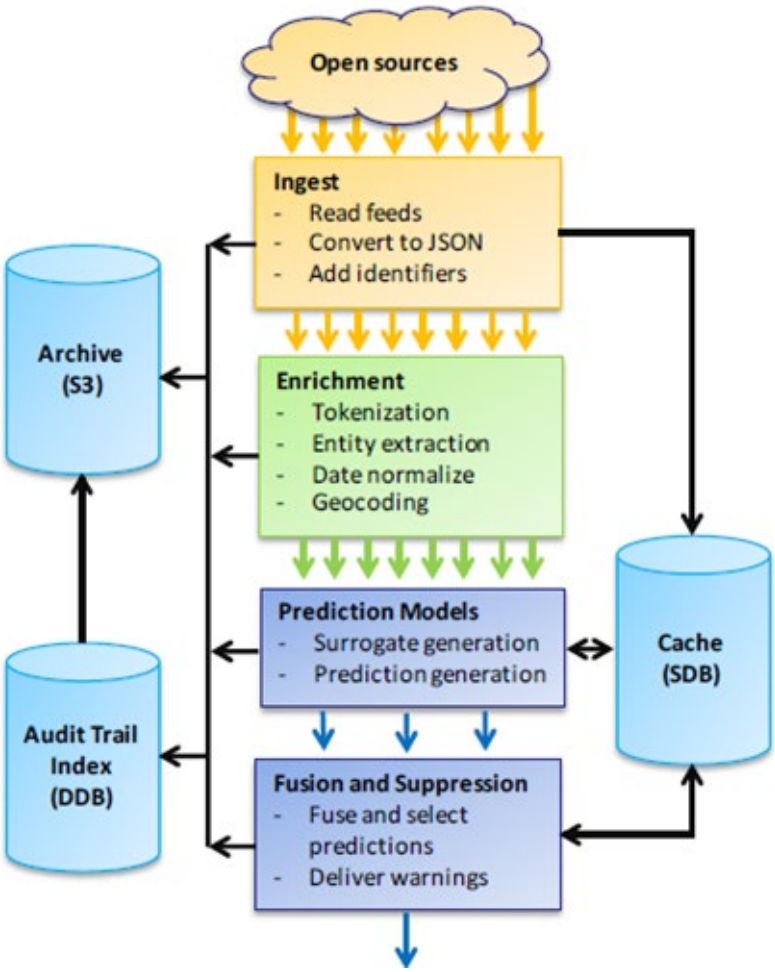

Figure 1-1: EMBERS Architecture

2.0 Glossary of Terms Development

Topically general data sources were identified to capture electronically published public data (articles, various media postings, etc.) as broadly as possible. The first problem to overcome was selecting data that is compatible with the objectives and resources of the project and without compromising the detectible signal from the data sources. Manageable development datasets were created by applying complex queries using an initial key word list or glossary on the huge data archives. This section describes the methods that are used to develop the key word glossary. These methods will be incorporated as part of the data ingestion for the prototype system.

Researchers created a glossary of terms to make a preliminary and high-level identification of the most relevant words or phrases pertaining to pit production mission at the SRS and other related nuclear events of interest. The glossary of terms serves three primary purposes:

1) to act as a set of seed terms to query open data sources and retrieve only articles or Tweets™[2] that are presumably related to some aspect of pit production at SRS, the nuclear weapons complex, or the nuclear fuel cycle (each in accordance with the events of interest outlined in (Danielson, Kail and Pike April 2020)),
2) to be used as a foundational set of terms against which articles and Tweets™ can be compared and from which a topical "similarity score" can be derived (rendering article text machine/model interpretable), and
3) to control the size of the data that is returned, thereby keeping data acquisition costs lower than a more general, universal query.

2.1 Events of Interest

Development of the SRPPF occurs through a series of events that culminate in the startup of the facility. This includes but is not limited to such events as funding decisions, process selection and development, facility design, construction, purchase of materials or equipment, hiring and concentration of specialized skills, visits or announcements of key political figures, and, after startup, waste and environmental releases. Details of the development of events of interest are published in (Danielson, Kail and Pike April 2020). To summarize, all nuclear related events are divided into four event domains: "Acquisition Events", "Political/Diplomatic Events", "Economic Events", and "Population/Personnel Events". Each event can be categorized to specific activity domains where the models and methods in this report focus on identifying the nuclear activity of "Fissile Core Fabrication" from all other activity. For a detailed description, see (Danielson, Kail and Pike April 2020).

It is important to note that the glossary development was implemented as a pre-requisite to any exploratory data analysis, which would eventually guide model development and/or dataset refinement that might be necessary based on the level of "signal" present in the data universe. Therefore, the intent was to create a generic glossary of terms consistent with the defined events of interest in addition to glossaries characteristic to event type. For the generic list, researchers sought to cast a broad net, capturing relevant "nuclear activities" terms while also capturing, for example, terms describing more tangentially occurring economic or political events (e.g., housing sales, business developments, and increased traffic in some geographic region, etc.) that could potentially act as indicators in forecasting the ultimate event of interest (i.e., pit production coming to the SRS). In other words, an overarching hypothesis in this early developmental stage was that a contextual weighted connectedness between all retrieved textual data could

[2] Tweet is a trademark of Twitter, Inc. or its affiliates.

eventually be identified and that embedded within would be various indicators, allowing for an inferential forecast.

2.2 Method

To capture relevant terms that should be included in the glossary of terms, a text frequency analysis was performed on nine openly released government documents (DOE May 2003) (DOE April 2020) (DOE April 2014) (DOE March 2015) (DOE March 2016) (DOE November 2017) (DOE October 2018) (DOE July 2019) (DOD 2020) and two DOE NNSA websites (DOE n.d.) (DOE n.d.) pertaining to the nuclear weapons complex and the plutonium pit production mission. The workflow for the text frequency analysis is illustrated in Figure 2-1, where each document's text was pre-processed (e.g., tokenization, lemmatization, and stop-word and numerical data removal), and the frequency of occurrence of 1-, 2-, and 3- grams was cumulatively updated. Experimentation was performed for higher degree n-grams, though no significant increase in performance (i.e., capture of relevant terms) was found.

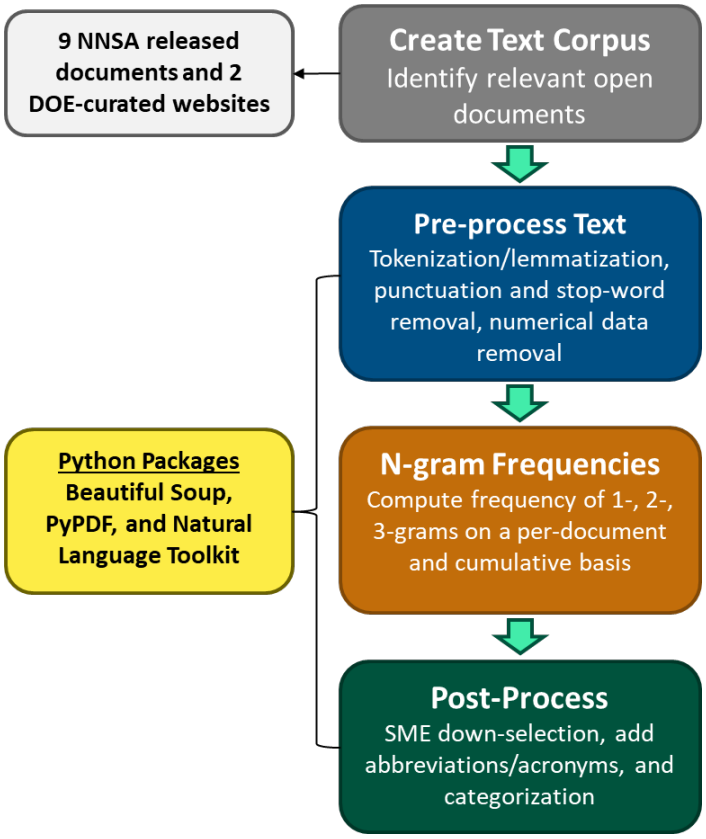

Figure 2-1: Text Frequency Analysis Workflow for Creating the Glossary of Terms

After processing each document, a cumulative text frequency distribution was formed containing approximately 613,000 n-grams. The top 75 most frequently occurring n-grams are shown in Figure 2-2, where the term *nuclear security* was the most frequently occurring term. This is perhaps unsurprising as many of the documents were released by the National Nuclear Security Administration, and therefore, this n-gram is frequently found in multiple contexts (i.e., on its own or as part of the NNSA four-word n-gram). Similarly, many of the n-grams shown are clearly only partial word phrases, though the full phrase can be easily inferred. While more sophisticated approaches exist, here, a more rudimentary manual down-

selection process was carried out by subject matter experts (SMEs) to create the final list of terms using the following workflow assumptions (in order).

1. Any n-gram occurring fewer than ten times is irrelevant.
2. Incoherent/incomplete terms can be removed unless further context can be readily inferred (e.g., *level exposure extending*, which occurred ten times in the composite list).
3. Duplicate terms can be removed (this includes partial n-grams for which the full word phrase can be easily inferred e.g., *nuclear security administration* → *national nuclear security administration*).
4. Terms that are either too specific and/or deemed unrelated to any event of interest can be removed (e.g., *flight test*).

The first assumption alone immediately decreases the size of the n-grams list to less than 10,000 terms and applying the next three assumptions resulted in a final list size of approximately 450 terms. After down-selection, the SMEs categorized n-grams into their most appropriate event definition category. The final lists of terms are shown in Appendix: Search and Glossary Terms.

This manual down-selection process allowed SMEs an immediate qualitative performance assessment of the approach (i.e., Are any obvious terms missing? Is there a good mixture of relevant entities, activities, and semi-technical terminology?, etc.). In general, the final list was evaluated as having produced a fairly comprehensive set of terms that capture many relevant aspects of pit production, the nuclear fuel cycle, and weapons development programs as outlined in the preliminary event definitions such that data queries could be formulated. While this approach is fairly straightforward and preliminary testing has proven some success (as shown in later sections of this report), a more automated and generalized algorithmic approach will be desirable as the current work evolves into more complex data environments (i.e., more obscure data environments).

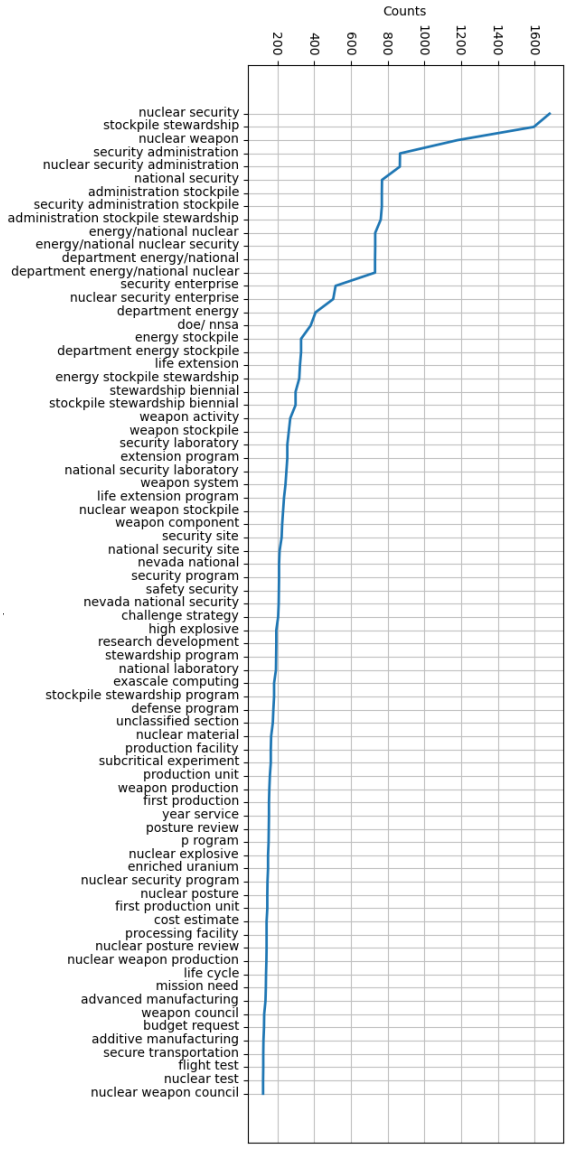

Figure 2-2: Top 75 Most Frequently Occurring n-grams From the Text Frequency Analysis Prior to Manual Down-Selection of Terms

3.0 Social Media – Twitter™ Dataset Development

Archived Twitter™ data that is readily available to the project team consists of a global decahose database (i.e., one in every ten Tweets™) that spans the period of August of 2014 to April of 2018. As mentioned in Section 1.0, the official announcement pertaining to the new pit production mission being assigned to the SRS came in May of 2018. Therefore, the Twitter™ archive captures a useful period for testing the prototype models' ability to forecast the SRPPF before it was announced. Initially, the level of signal (i.e., Tweets™ concerning events of interest) within the archive was unknown. Figure 3-1 illustrates a high-level workflow that was used for exploratory data analysis (i.e., an assessment of the available signal) and prototype model development. The following subsections provide specific details and results from this workflow.

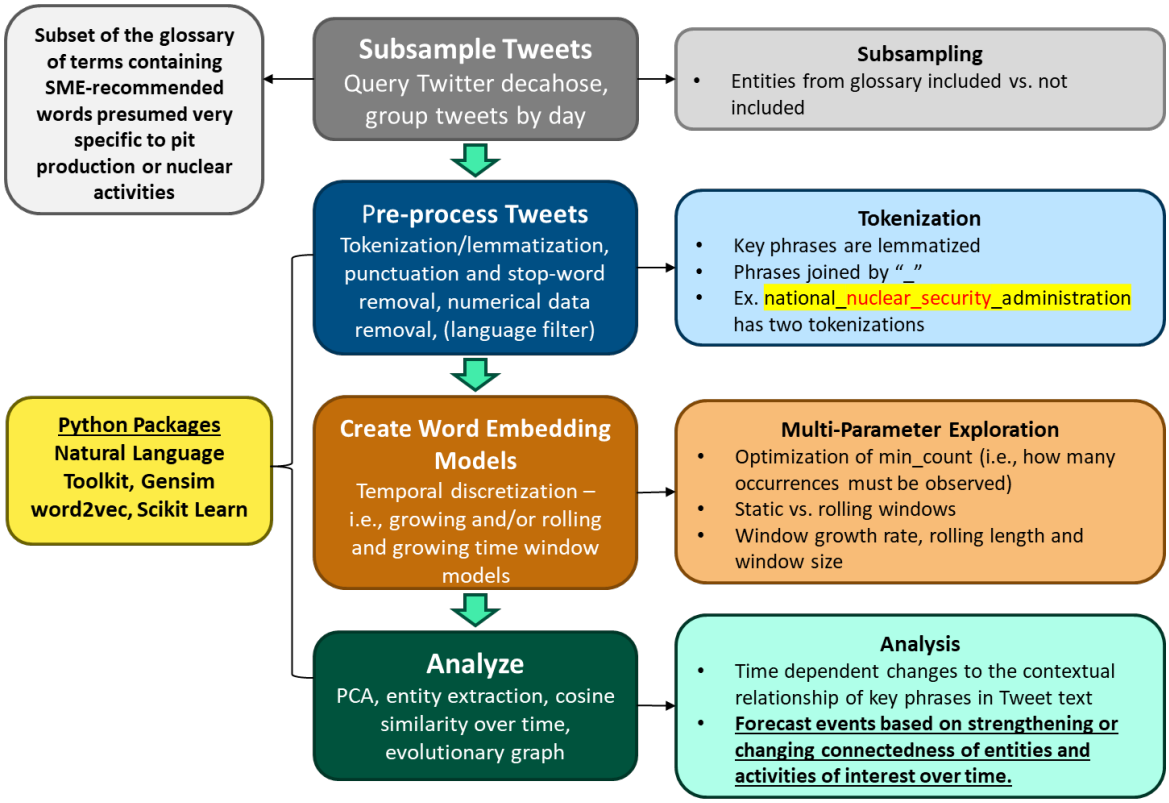

Figure 3-1: High-Level Workflow for Preliminary Model Development Using the Twitter™ Dataset

3.1 Data Sampling and Preparation

Given that Twitter™ imposes character limitations, Tweet™ text is generally very limited in contextual information and tends to be a noisy data environment. Therefore, to reduce the level of noise in the dataset, a targeted sampling of Tweets™ is performed by querying the decahose database with a subset of 89 terms (shown in Table 3-1) down-selected from the complete glossary of key terms in Appendix: Search and Glossary Terms. This subset was created by SMEs and places emphasis on the ability to extract only Tweets™ that contained explicitly mentioned terms that would presumably be very likely to correspond to some aspect that leads to the SRPPF project, a related process in the nuclear fuel cycle, or specific events or entities closely related to either. Also note that, as a subset of terms to generate an easily manageable

development dataset from the billions of tweets in the archive, the subset contains only part of many equivalent terms provided in the full glossary.

Table 3-1: List of Terms Used to Sample the Twitter™ Decahose Database

carlsbad site	nuclear component	plutonium processing
congressional defense committee	nuclear deterrent	plutonium processing facility
defense nuclear nonproliferation	nuclear explosive	Pu
defense secretary	nuclear facility	radioactive liquid waste
department of defense	nuclear incident	radioactive waste
department of energy	nuclear material	rocky flats
dod	nuclear nonproliferation	sandia national laboratory
doe	nuclear posture	savannah river site
doe facility	nuclear regulatory commission	secretary of defense
doe site	nuclear security	secretary of energy
enriched uranium	nuclear security enterprise	spent nuclear material
fissile component	nuclear stockpile	stockpile stewardship
fissile material	nuclear warhead	surplus plutonium
fission event	nuclear weapon	transuranic waste
hanford site	nuclear weapon component	tritium enterprise
lawrence livermore national laboratory	nuclear weapon council	tru waste
los alamos	nuclear weapon facility	uranium processing facility
los alamos national laboratory	nuclear weapon program	war reserve plutonium
low level waste disposal facility	pantex plant	warhead life extension
national defense authorization act	pantex site	weapon assembly
national nuclear security administration	pit disassembly	weapon complex
national security	pit facility	weapon component
national security complex	pit manufacturing	weapon dismantlement disposition
national security laboratory	pit production	weapon infrastructure
nevada national security site	pit production environmental impact statement	weapon modernization
nevada test site	plutonium	weapon production site
nnsa	plutonium disposition	weapon stockpile
nnsa infrastructure	plutonium facility	weapon system
nnsa site	plutonium operation	weapons grade plutonium
nonproliferation treaty	plutonium pit	

The sampling query retrieved slightly greater than 2.9 million Tweets™ across the three-and-a-half-year period and a Tweet™ count from each month is shown in Figure 3-2. The disproportionate spike around January 2017 is a prime example of the need to use a targeted sampling approach. During that month, there was a large quantity of Tweet™ traffic discussing weapons testing in North Korea and correspondingly, the Department of Defense, and the incipient stages of high-profile national security investigations within the Executive Branch of the US Government. In other words, none of these events are directly related to

the targeted event of interest, though the nature of the Tweets™ are presumed to be generally contextually similar to the types of Tweets™ that would be directly beneficial, and therefore they are not discarded.

In the sampling process, each Tweet™ was extracted in json format and contained a datetime timestamp, the Tweet™ body, and geolocation information, if any is available. To prepare Tweet™ text for machine learning models, each Tweet™'s text was first tokenized and lemmatized in all lower-case letters using the Python package Natural Language Toolkit. Subsequently, stop words and numerical data were removed as they generally provide minimal additional contextual information but can create "noise" in word textual analysis. The vast majority of Tweets™ are returned in the English language and therefore, no language filtration was deemed necessary.

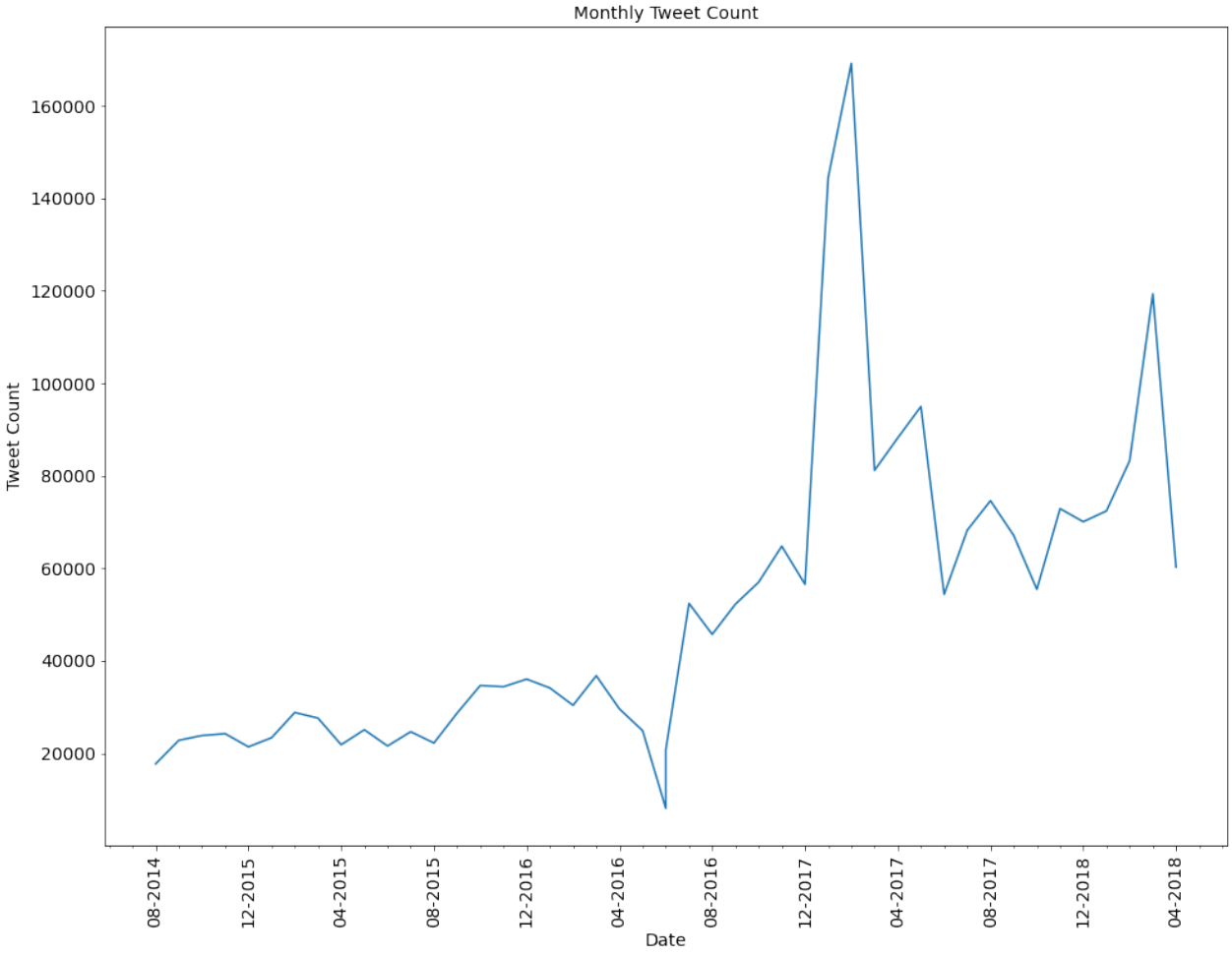

Figure 3-2: Monthly Tweet™ Count Returned by the Sampling Query

3.2 Word Embedding Models

A word embedding model is a machine learning technique that allows one to predict context given a word/set of words, or predict a word/set of words, given context. This capability is derived from the vectorization of words contained in a textual data corpus, whereby the similarity of words can be obtained by comparing the word vectors (e.g., by computing the cosine angle between two or more word vectors). Thus, this technique is ideally suited for analyzing Twitter™ datasets as it will provide the capability to analyze a Tweet™, which perhaps does not explicitly mention any specific terms from the glossary, but

contains words or phrases that have high similarity to words in the glossary, thereby allowing for prediction of contextual similarity to an area of interest. Note that all Tweets™ here will contain some number of terms from the glossary subset. Therefore, the subsampled dataset is for creating a trained word embedding model, where a well-trained model can then be applied to a global unfiltered dataset and contextual similarity can be obtained.

In addition to this predictive capability, one can analyze the evolution of word embedding models over time, whereby a changing context of words within the corpus (i.e., through the collection of more Tweets™ through time) might indicate a more broad change in an overarching paradigm and allow for a probabilistic forecast of specific events. In FY20, a thorough exploration of this potential was carried out on the Twitter™ dataset to provide a proof-of-concept that will enable further prototype model development.

3.2.1 Exploratory Data Analysis of Word Embedding Models

After textual pre-processing, the entire corpus of 2.9 million Tweets™ was used to train a word embedding model using Word2Vec as implemented in the Python package Gensim's continuous bag of words training algorithm. This initial testing served to provide evidence that the dataset is large enough, and dense enough, to train a model that can predict contextually similar terms to the targeted pit production at the SRS. A noteworthy capability of word embedding models is the ability to perform "word arithmetic", which can be illustrated using an example of *King – Man + Woman = Queen*, where each word is represented by a word vector in a trained word embedding model. This property was leveraged to explore the quality of the trained word embedding model. Figure 3-3 shows a list of the trained model's top ten most similar words to *pit* (i.e., shorthand for plutonium pit), *production*, and *pit + production*. Notably, and as expected, the most similar words for *pit + production* are different than those for the two words separately. The results indicate that the word embedding model has some level of capability of identifying that, here, these terms are to be understood as related to nuclear activities. This is demonstrated, for example, by the high similarity of terms like *tritium*, *reprocessing*, and *strontium*. Interestingly, *pit + production* returns *mox* as the most similar term – a promising result in that the proposed SRPPF will utilize infrastructural resources in place from the abandoned Mixed Oxide Fuel (MOX) project at the SRS.

```
In [14]: word2vec.wv.most_similar(positive=['pit','production'])

Out[14]: [('mox', 0.6826311349868774),
          ('mountain', 0.6515145897865295),
          ('decaying', 0.6500195264816284),
          ('producing', 0.6485304236412048),
          ('tritium', 0.6459552049636841),
          ('output', 0.6454975605010986),
          ('cobalt', 0.6450344324111938),
          ('restarting', 0.6286609172821045),
          ('disposing', 0.627967357635498),
          ('repository', 0.6238623857498169)]

In [15]: word2vec.wv.most_similar(positive=['pit'])

Out[15]: [('mountain', 0.6103617548942566),
          ('saturn', 0.5715295672416687),
          ('decaying', 0.5650793313980103),
          ('cobalt', 0.5465441942214966),
          ('truck', 0.5420590043067932),
          ('strontium', 0.5395830273628235),
          ('powered', 0.5318939089775085),
          ('tritium', 0.5316237211227417),
          ('beryllium', 0.5258057117462158),
          ('canyon', 0.5235276222229004)]

In [16]: word2vec.wv.most_similar(positive=['production'])

Out[16]: [('output', 0.6688458919525146),
          ('mox', 0.6672303676605225),
          ('producing', 0.6570849418640137),
          ('restarts', 0.6445433497428894),
          ('disposal', 0.6387706995010376),
          ('disposing', 0.6335718035697937),
          ('reprocessing', 0.6244850158691406),
          ('repository', 0.6119323968887329),
          ('smr', 0.611620306968689),
          ('yongbyon', 0.6105226874351501)]
```

Figure 3-3: Top 10 Most Similar Words to *pit* + *production*, *pit*, and *production* From the Trained Word Embedding Model, With Cosine Similarity Scores Listed

As a refinement to this initial testing, an additional pre-processing step was implemented where the Tweets™ were tokenized by having multi-word glossary terms joined by an underscore (e.g., *pit_production*). In this tokenization process, if a Tweet™ contains, for example, national nuclear security administration (which contains two key word groups from the glossary: national nuclear security administration and nuclear security) two tokenized Tweets™ are returned: one containing *national_nuclear_security_administration*, and the other containing only *nuclear_security*. Note that the same lemmatization process is applied to the terms from the glossary for consistency and comparison. Such a step is a straightforward approach to implementing a semi-Phrase2Vec type model whereby multi-word n-grams are identified as a single word. After this re-tokenization process, the word embedding model was re-trained and word similarities were again explored to check for improvement. Figure 3-4 shows the word similarities for *pit* + *production*, *pit*, and *production* after re-tokenization. No noteworthy performance increase was gained when exploring cosine similarity lists of the words separately, though there was also not a perceived performance decrease. However, Figure 3-5 shows the word similarities to *pit_production* which shows a performance increase by capturing entities such as *rocky_flats*, which is a former site for pit production activities. Based on this result, among others, the re-tokenized method was selected for use.

```
In [12]: word2vec.wv.most_similar(positive=['pit','production'])

Out[12]: [('liquid', 0.6805588006973267),
          ('americium', 0.6630449295043945),
          ('thermal', 0.6373246908187866),
          ('output', 0.6371421217918396),
          ('dounreay', 0.6362630128860474),
          ('generator', 0.6354036331176758),
          ('pellet', 0.6287705898284912),
          ('reprocessing', 0.6261838674545288),
          ('strontium', 0.6203876733779907),
          ('extraction', 0.6187630891799927)]

In [14]: word2vec.wv.most_similar(positive=['pit'])

Out[14]: [('stricken', 0.5667874813079834),
          ('laced', 0.5575540661811829),
          ('himalaya', 0.5556865930557251),
          ('paradise', 0.5520955920219421),
          ('spray', 0.5518181920051575),
          ('napalm', 0.5491824746131897),
          ('pellet', 0.5404135584831238),
          ('melted', 0.5367121696472168),
          ('cargo', 0.5327005982398987),
          ('radium', 0.5302586555480957)]

In [15]: word2vec.wv.most_similar(positive=['production'])

Out[15]: [('output', 0.6829184293746948),
          ('reprocessing', 0.6631824374198914),
          ('producing', 0.6487859487533569),
          ('thermal', 0.6426633596420288),
          ('breeder', 0.6315975189208984),
          ('enrichment', 0.6302756071090698),
          ('mox', 0.6210713386535645),
          ('extraction', 0.6199362277984619),
          ('extracting', 0.6125844717025757),
          ('liquid', 0.6109449863433838)]
```

Figure 3-4: Most Similar Words to *pit* + *production*, *pit*, and *production* After Re-tokenization

```
In [13]: word2vec.wv.most_similar(positive=['pit_production'])

Out[13]: [('tritium', 0.7282271981239319),
          ('nuclearwaste', 0.7112715244293213),
          ('rockyflats', 0.7103050351142883),
          ('nuclearpower', 0.7093418836593628),
          ('mox', 0.7076944708824158),
          ('repository', 0.6868902444839478),
          ('thorium', 0.6862359046936035),
          ('reprocessing', 0.6756302714347839),
          ('extraction', 0.6662992238998413),
          ('paducah', 0.6605997085571289)]
```

Figure 3-5: Most Similar Words to *pit_production* After Re-tokenization

To explore the relationship between all words in the Twitter™ database, as represented by the word embedding model, a principal component analysis (PCA) was performed, whereby the 100-dimensional word vectors were collapsed to 2 dimensions based on the most highly contributing vector components. Figure 3-6 shows the PCA with only the keywords from the glossary of terms. Notably, at this high-level view, a tight cluster of words is centered around the point (0, 0) and terms are more scattered moving away. Further away from the origin, high-level governmental agencies (e.g., department of energy) and more broad terms (e.g., nuclear security) are found as these terms are likely to appear in broad contexts. Figure 3-7 shows a zoomed in PCA plot moving closer to the origin where more specific entities, such as national laboratories and DOE sites are found. Again, these terms are likely to be mentioned in Tweets™ within broad contexts for things such as research accomplishments in press releases, but they also occur frequently

with nuclear activities-type words. Finally, Figure 3-8 shows a zoomed in plot of the PCA around the origin where much more specific terms related to nuclear materials and weapons production are found.

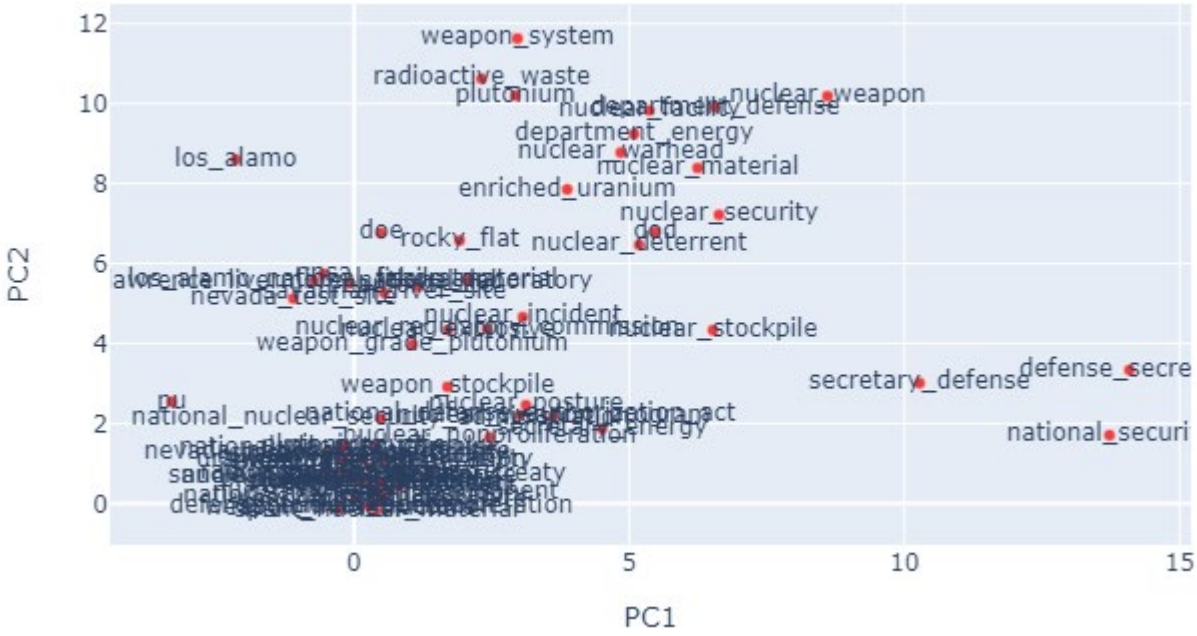

Figure 3-6: Principal Component Analysis Showing the Relationship Between All Word Vectors in the Trained Word Embedding Model's Vocabulary

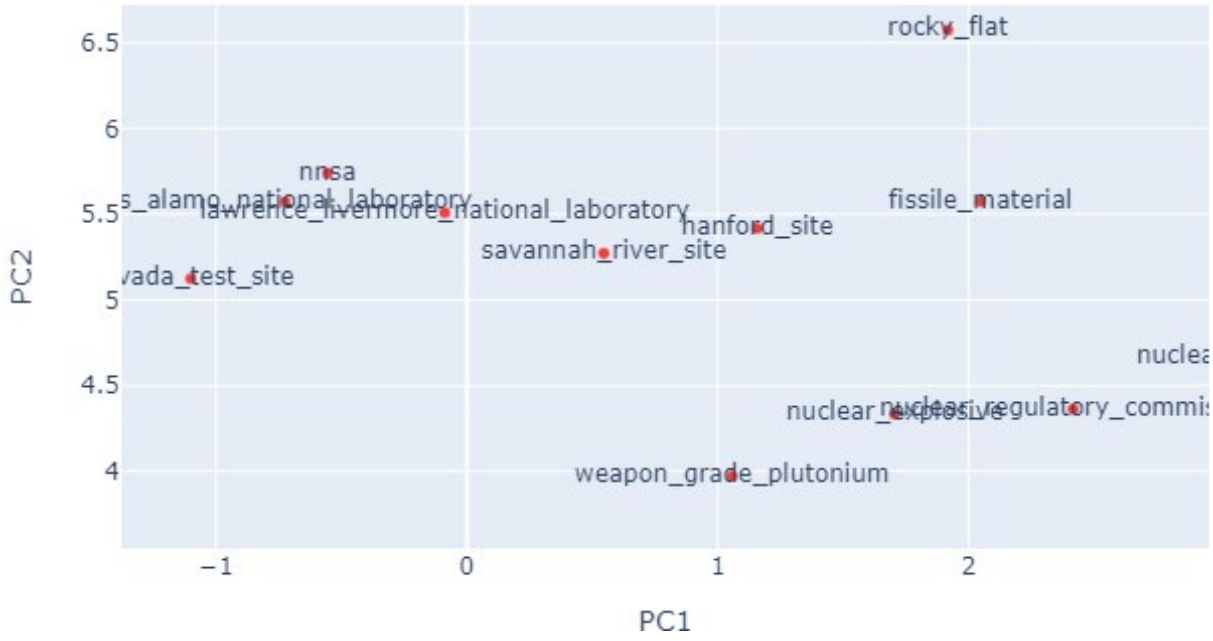

Figure 3-7: Zoomed in Snapshot of the Principal Component Analysis Just Outside the Tight Cluster of Words

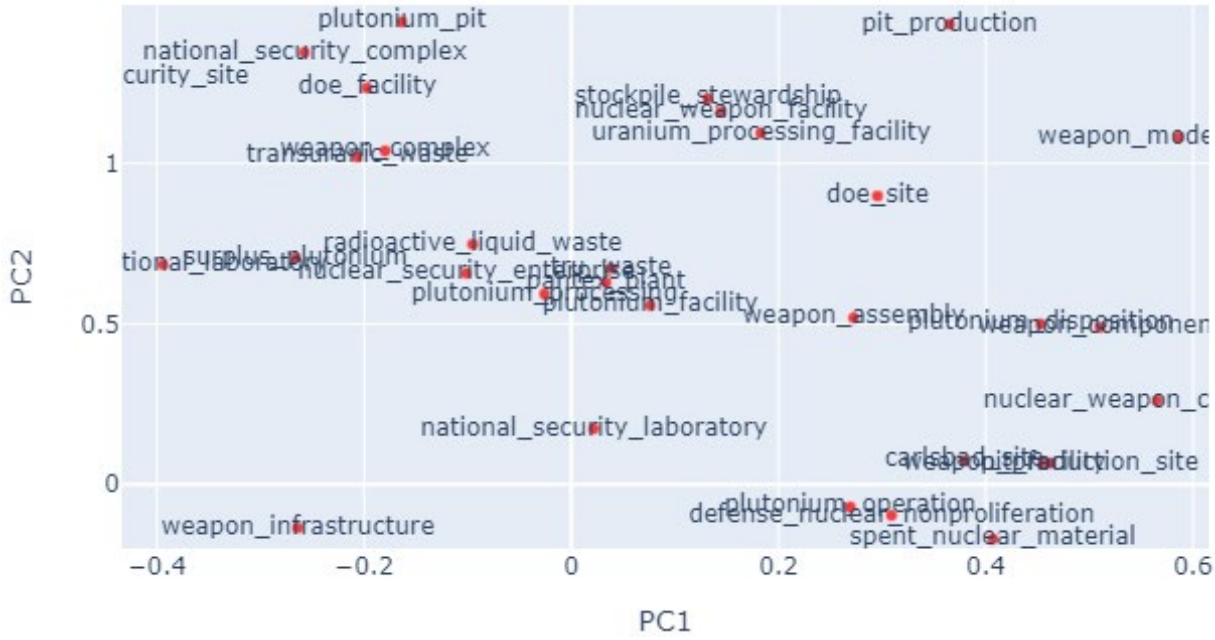

Figure 3-8: Zoomed in Snapshot of the Principal Component Analysis on the Tight Cluster of Words

3.3 Time Dependent Word Embedding Models

As mentioned previously, time-dependent changes in word embedding models can be used to extract events that are indicated by shifts in the contextual models. Here, two types of time dependent word embedding models were developed for use in the prototype forecasting system: 1) a static, continuously growing time window and 2) a constant sized, rolling window. The following subsections present the key results from these two algorithms.

3.3.1 Static Continuously Growing Time Window Model

The algorithm for the static, continuously growing time window model is outlined as follows:

1. Select a growth rate, *r*, which is the length of time represented each time new Tweets™ are added and a new word embedding model is trained.
   a. Note that the number of Tweets™ added at each new time interval is not constant.
2. Initialize data set at time $t = 0$ with a window of size *r*.
3. Train a word embedding model.
4. At time $t = r$, add all Tweets™ from $t = r$ to $t = 2r$.
5. Train new word embedding model.
6. Repeat step 4 until models exist across all time, where the window size grows by *1r* at each new interval.

Here, *r* has been set to 1 month, though this parameter may need to be optimized for the prototype system. The vocabulary size over time for the word embedding models is shown in Figure 3-9 and ranges from approximately 7,700 to 110,000 words.

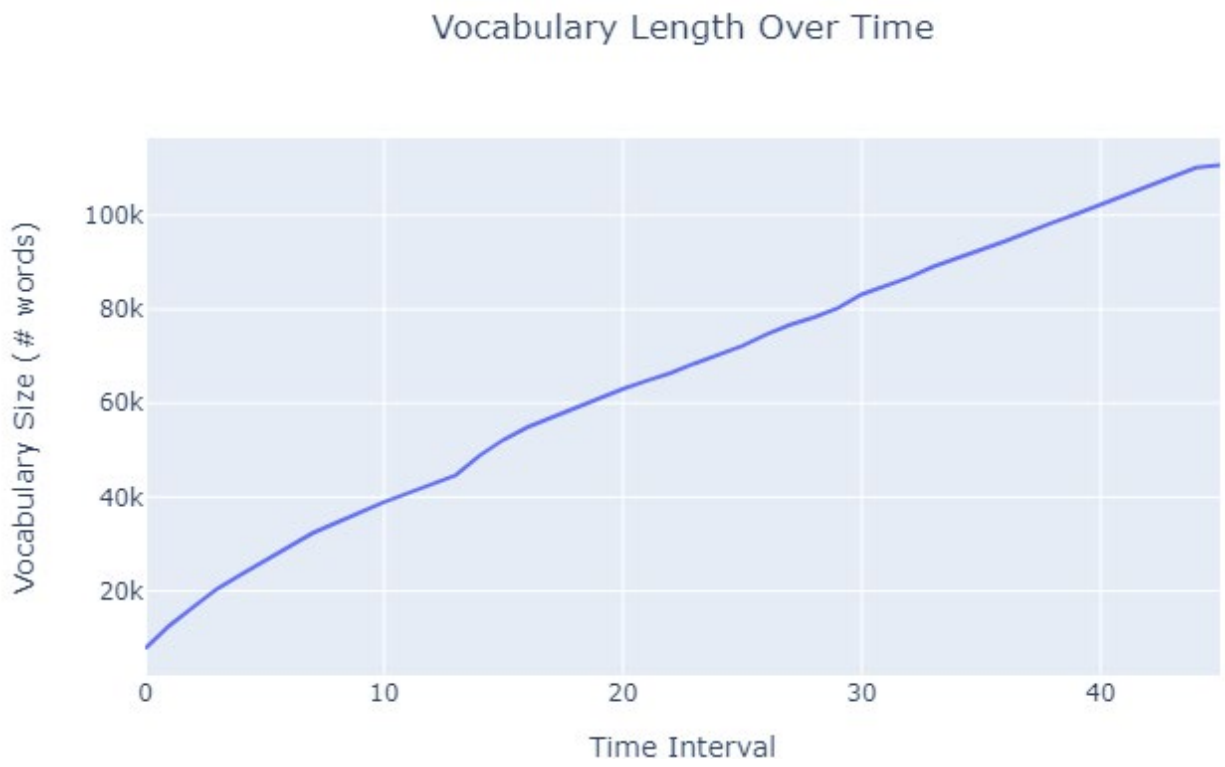

Figure 3-9: Vocabulary Length of Word Embedding Models Over Time and the Data Window Grows

To explore how the word embedding models change as they grow and identify contextual shifts of key terms, the similarity rank over time is explored. To compute the similarity rank, the most similar words in the vocabulary to *pit_production* are scanned until *savannah_river_site* is found. The index in the list is normalized against the vocabulary size, as each word embedding model's vocabulary is a different length. By exploring the word embedding models in this way, multiple path lengths can potentially represent the shortest distance between the two phrases. For example, if *savannah_river_site* is not the nearest neighbor (NN) (e.g., the most similar word) to *pit_production*, it could potentially be a NN to the NN of *pit_production*. A plot for this example is shown in Figure 3-10. Notably, *savannah_river_site* is identified as the most similar word to *pit_production* as early as March of 2016 – two years prior to the official announcement. Figure 3-11 shows this analysis again where the similarity rank of *pit_production* with respect to *savannah_river_site* is explored. Notably, while *pit_production* is never the most similar word to *savannah_river_site*, it is most frequently found within the top 0.1% to 1% of the vocabulary. Finally, Figure 3-12 shows the similarity rank over time between *doe_site* and *savannah_river_site*. This example demonstrates a case where a term can be more closely related to the seed phrase (here, *doe_site*) through the NN's most similar words list. This is shown by the yellow marker at month 40, where *savannah_river_site* was the 3rd most similar word to *doe_site* in the vocabulary and *hanford_site* was the most similar word. However, *savannah_river_site* was the most similar word to *hanford_site*. In other words, the connection was made one step earlier. While this example does not demonstrate a major improvement in the proximity, the potential exists to significantly decrease the path length between two terms using this approach of scanning all path lengths.

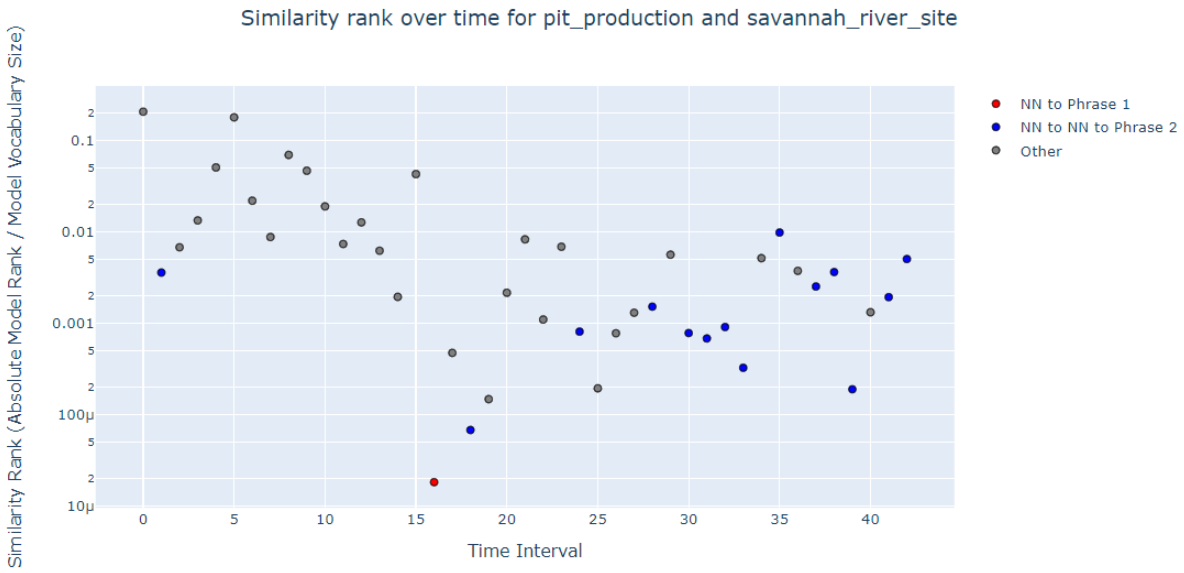

Figure 3-10: Similarity Rank of *savannah_river_site* Over Time Relative to *pit_production*

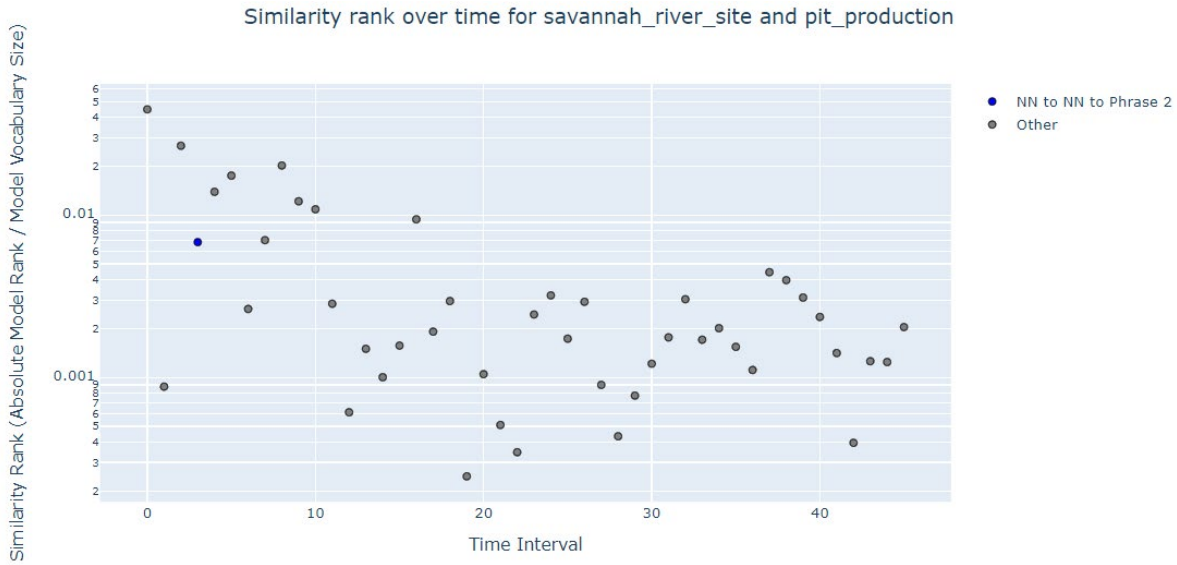

Figure 3-11: Similarity Rank of *pit_production* Over Time Relative to *savannah_river_site*

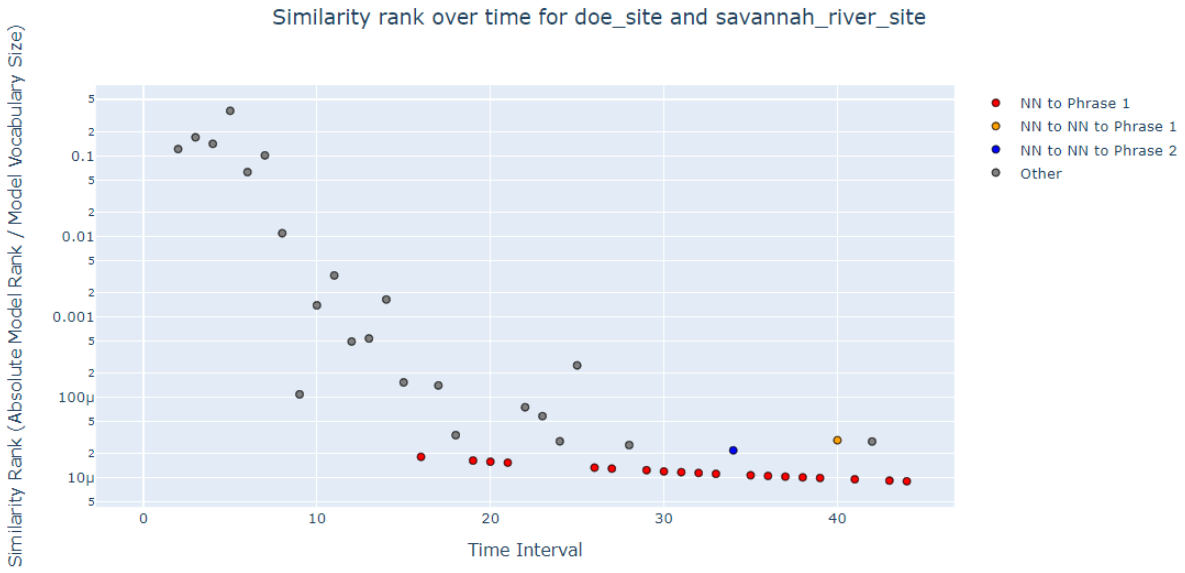

Figure 3-12: Similarity Rank of *doe_site* Over Time Relative to *savannah_river_site*

In addition to the similarity rank over time, the top N most similar words to a keyword can be plotted over time to identify new entities or terms that may appear. Figure 3-13 shows a plot of the top 50 most similar words to *pit_production* for the time-dependent word embedding models. Note that *savannah_river_site* appears in the top 50 most similar words as early as September of 2014. However, note that the earlier word embedding models have a relatively small vocabulary when compared to the later and larger word embedding models. This explains the apparent convergence in the cosine similarity plot. Notably, several key entities in the DOE nuclear weapons complex, including *rocky_flats*, *lanl*, and *savannah_river_site* are frequently in the top 50 most similar words to *pit_production* over time.

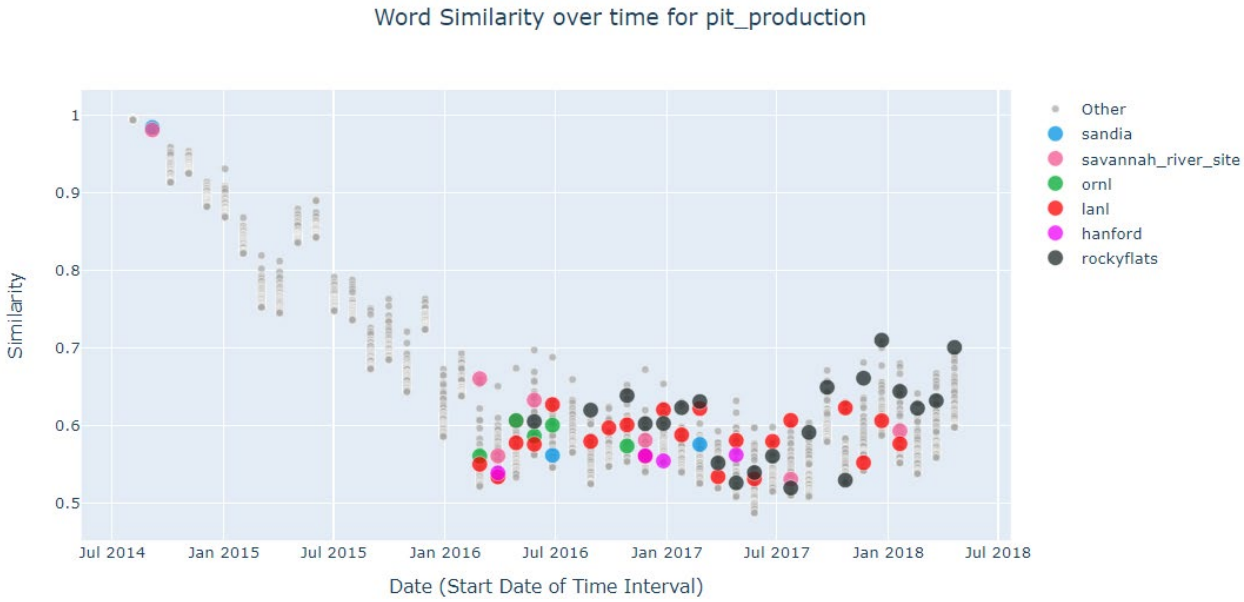

Figure 3-13: Top 50 Most Similar Words Over Time to *pit_production* With Several Key Entities Highlighted

These results demonstrate that the Twitter™ data allows for entities to be adequately captured by word embedding models. However, most of the entities that have been captured were included in the original glossary of terms, and therefore the database query returns any Tweet™ containing an entity from the glossary. In a production application, all entities may not be known. Therefore, to test if the word similarity between entities is a result of the original query, subsampling of the dataset was performed with all Tweets™ that only contain an entity keyword removed (i.e., if a non-entity keyword and an entity keyword appear in a Tweet™, the Tweet™ is kept). This reduced the number of Tweets™ to approximately 1.2 million. The similarity rank over time for *pit_production* and *savannah_river_site* and the top 50 most similar words to *pit_production* over time on the sub-sampled Tweets™ demonstrate a similar performance as the full dataset, where *savannah_river_site* is the most similar word to *pit_production* as early as October of 2016 and several key entities that are closely related to pit production (in reality) are still identified. This is a key result toward the proof-of-concept that such a technique is applicable even if all entities are not identified prior to querying a dataset.

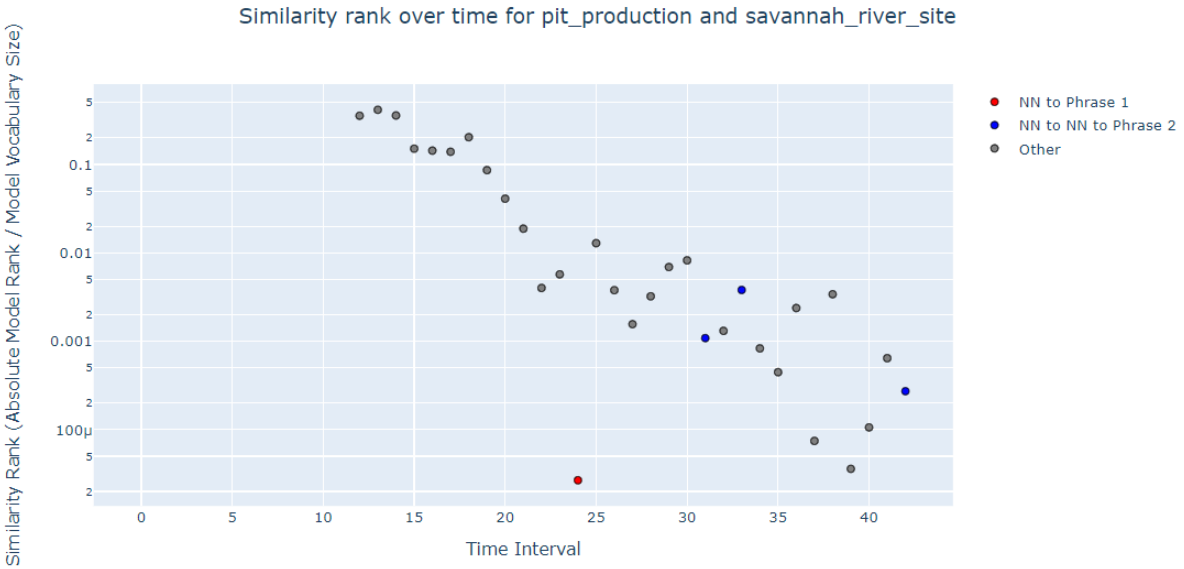

Figure 3-14: Similarity Rank Over Time Between *pit_production* and *savannah_river_site* When the Twitter™ Dataset is Sampled Without Entity Keywords

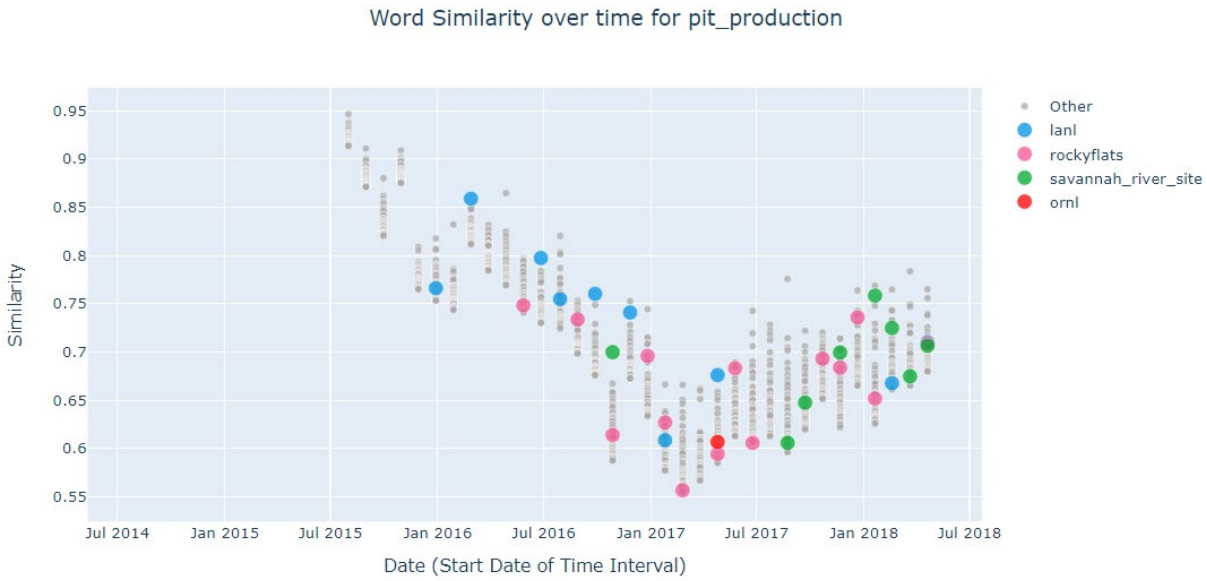

Figure 3-15: Top 50 Most Similar Words to *pit_production* Over Time When the Twitter™ Dataset is Sampled Without Entity Keywords
(Note: Several key entities are highlighted.)

3.3.2 Constant Sized, Rolling Time Window Model

The algorithm for the constant sized, rolling window model is outlined as follows:

1. Select a window size, *w*, representing a length of time across which Tweets™ are held and a rolling length of *r* representing a length of time at which Tweets™ are dropped each time rolling occurs.
2. Initialize data set at time $t = 0$ with a window size *w*.
3. Train a word embedding model.
4. Shift window of size *w* by *r* adding Tweets™ at the end and removing Tweets™ from the first *r* length of time.
5. Train a new word embedding model.
6. Repeat Step 4 until the end of data is reached.

Two different parameterizations have been tested: a 365-day window size with a rolling length of 30 days and a 730-day window size with a rolling length of 30 days. Again, the rolling length and the window size are two parameters that will need to be tailored to the specific use in the demonstration prototype. The vocabulary length over time of the two models are shown in Figure 3-16 and Figure 3-17. Notably, the vocabulary sizes in this model is more constant over time, varying by only about 10-15%.

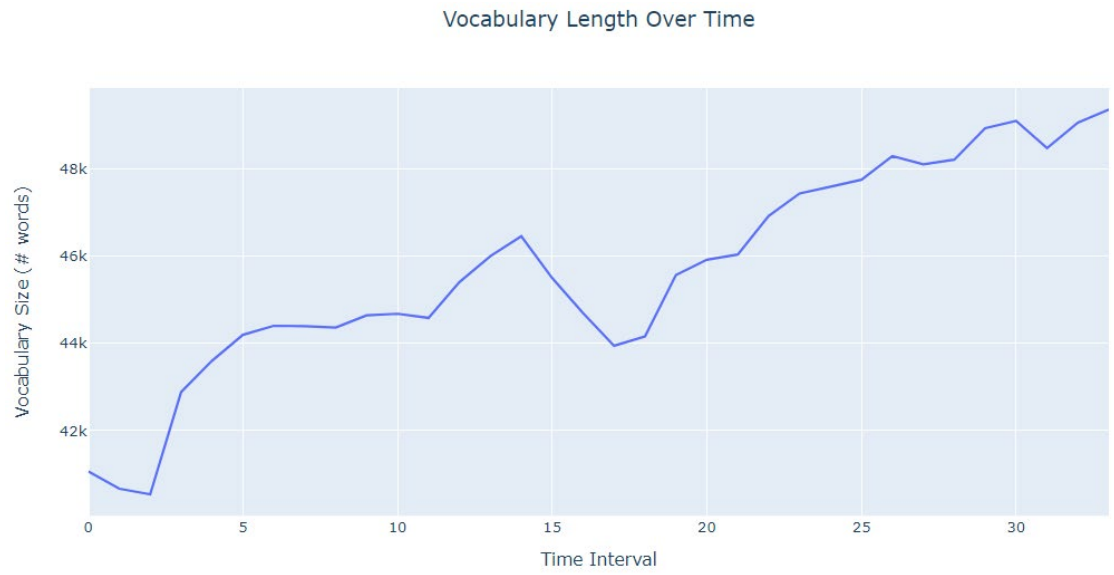

Figure 3-16: Vocabulary Size Over Time for a 365-Day Rolling Window and a 30-Day Rolling Length

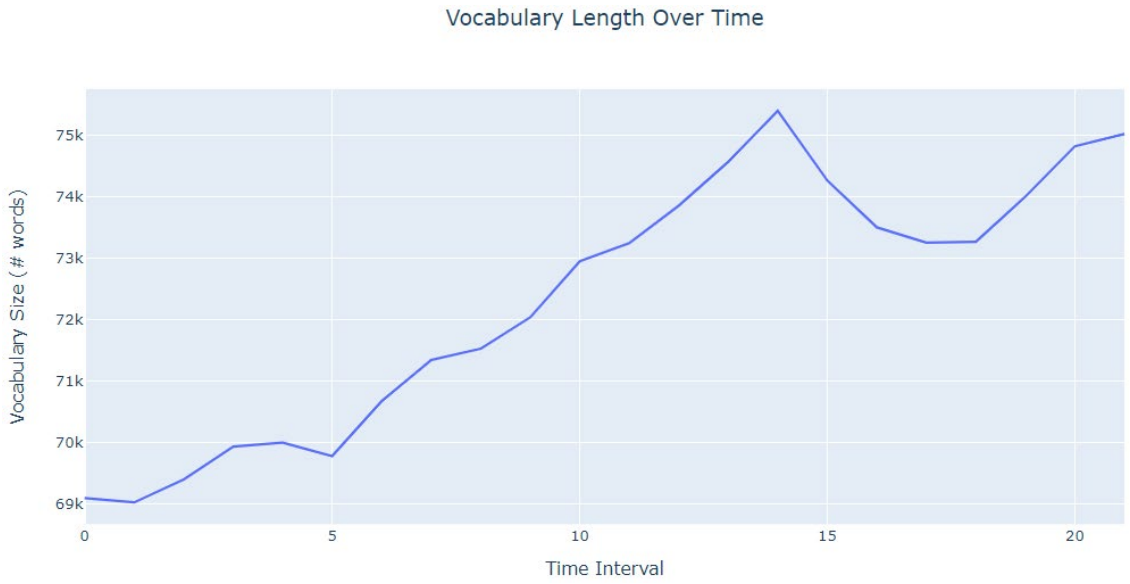

Figure 3-17: Vocabulary Size Over Time for a 730-day Rolling Window and a 30-Day Rolling Length

The similarity rank over time and the top 50 most similar words over time are shown for the two differently parameterized models in Figure 3-18 through Figure 3-21. Notably, *savannah_river_site* is never identified as the most similar word to *pit_production* for either parameterization. Additionally, the term *pit_production* falls out of the vocabulary for several months with a window size of 365 days. While some of the key entities are still identified over time in the top 50 most similar words, the frequency of all entities being recognized as similar is much lower. This may indicate that the models are not as adequately trained as the cumulative growing model because data is dropped out of the word embedding models over time.

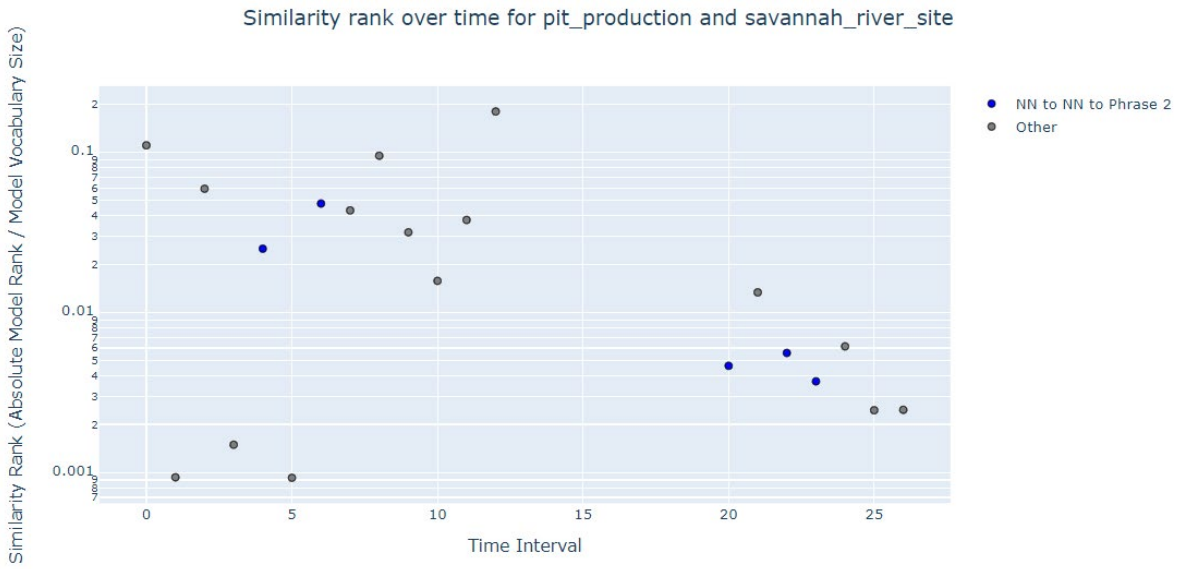

Figure 3-18: Similarity Rank Over Time for *pit production* and *Savannah River Site* for a 365-Day Rolling Window and a 30-Day Rolling Length

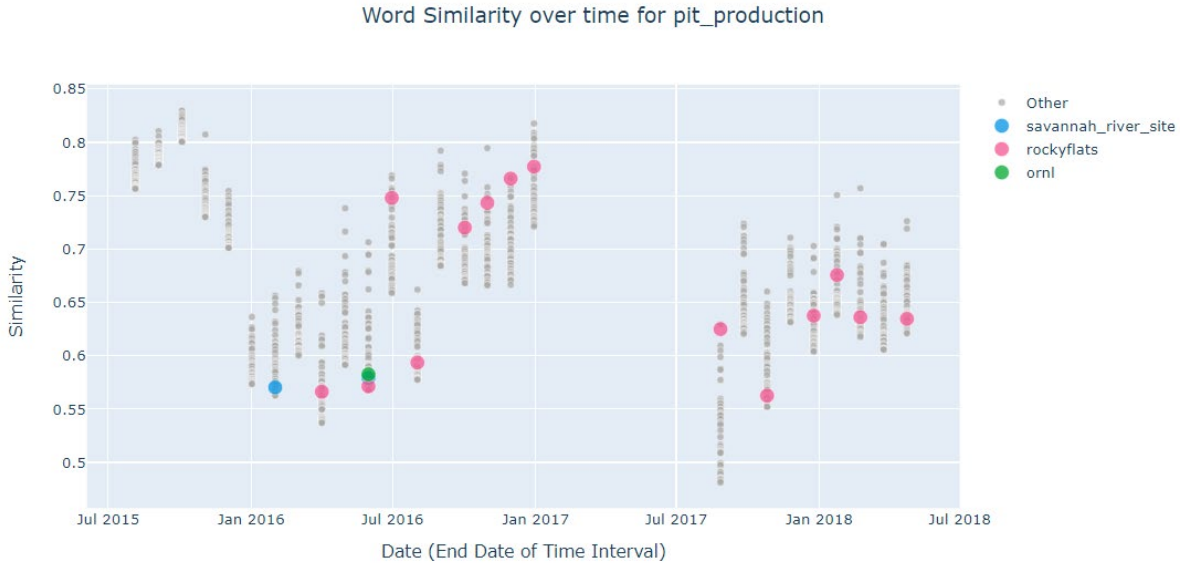

Figure 3-19: Top 50 Most Similar Words to *pit production* Over Time for a 365-Day Rolling Window and a 30-Day Rolling Length

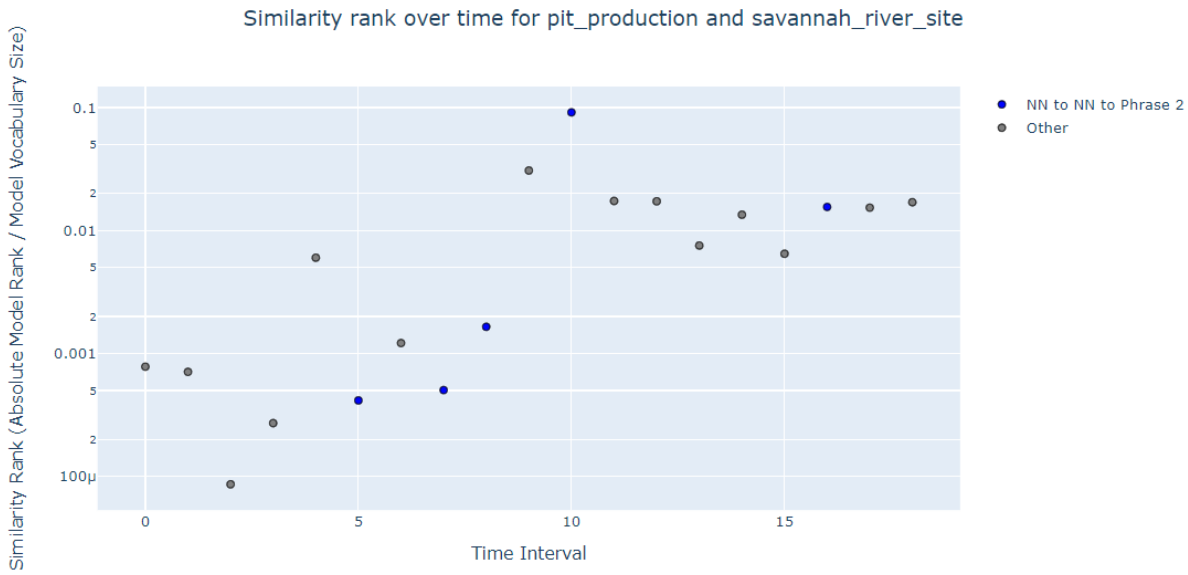

Figure 3-20: Similarity Rank Over Time for *pit production* and *Savannah River Site* for a 730-Day Rolling Window and a 30-Day Rolling Length

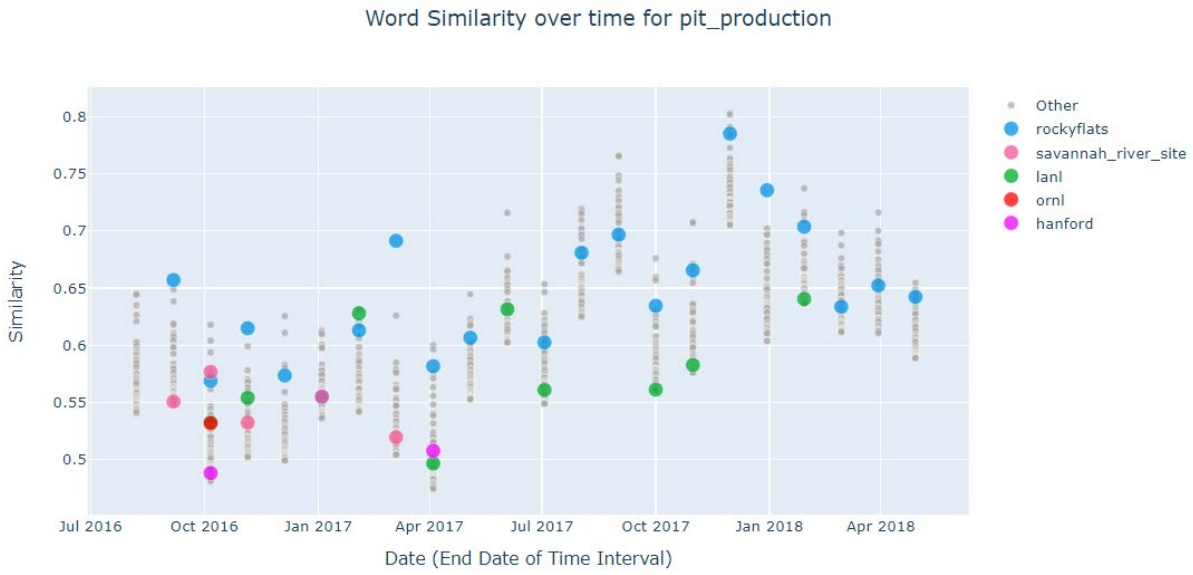

Figure 3-21: Top 50 Most Similar Words to *pit production* Over Time for a 730-Day Rolling Window and a 30-Day Rolling Length

4.0 News Article Aggregator Data Source – Webhose Ltd.

The following sections provide detailed descriptions of the modeling pipeline as developed for the news article data along with the results to characterize the effectiveness and provide insight into the working of each facet of the multi-step modeling pipeline under development. The pipeline is consistent with the planned prototype system architecture shown in Section 1.3.

Figure 4-1 shows the multi-step modeling that follows after data ingestion for an event detection system. The following paragraphs provide a description of each block.

Figure 4-1: Event Detection Model Pipeline for Anomaly Detection

Create Word Embedding: Given a set of unprocessed news articles, S, and a set of seed-phrases of interest, P, provided by a domain expert, word embedding models are trained on textual data to enable inference of the content of each article and phrase. The effect of word embedding models allow similarity comparisons between a news article and any given phrase. From these models, researchers developed a methodology that is able to transfer documents (i.e., news articles) and phrases into a common space to enable distance calculations between vectorized textual data.

Search: Similarity calculations between phrases and documents allow identification of the most similar documents with respect to specific phrases. A multi-stage ranking procedure based partially on the latent document and phrase embeddings obtained from the word-embedding pipeline has been developed. The ranking procedure reveals a subset of relevant documents, D, from the entire universe of news articles, S, such that $|D|<<|S|$ where $|A|$ indicates the cardinality of set A.

Entity Extraction: Subsequent steps all focus on the set of relevant documents, D. For each document, important themes and topics must be identified in a manner more overt than that provided by latent document embeddings in the first step of the pipeline. Hence, a key step in the pipeline is to extract entities of interest from each document in D. The entities capture agency names including multiple variations of the name, facility identities/names, locations, or dates of interest where each entity is tagged with their corresponding types.

Anomaly Detection / Event Characterization: Documents with entities enriched are employed for creation of a weighted temporal heterogeneous entity graph, G. This graph evolves over time and reveals the evolution of the relationships between entities. The evolution reflects the operations of various entities of interest and provides the following information:

1. Characterization of the evolution of entities of interest.
2. Detection (or forecast) of potentially anomalous relationships between entities in the graph, thereby providing experts a focused sub-set of documents D' such that $|D'|<<|D|$ for further investigation.

4.1 Data Sampling and Preparation

News articles ingested into the pipeline were obtained from Webhose Ltd. (https://webhose.io/). Webhose was chosen because they had the largest historical (i.e., oldest) dataset and provided article text along with metadata. The researchers initially used a query containing the full glossary of terms defined in section 2.0. An initial analysis of the results returned by this query showed that many results were not related to nuclear activity. Figure 4-2 shows a simple word embedding analysis that of the queried documents, the ones that contain the terms nuclear or plutonium (red dots) are clustered and the blue dots are further out and less relevant to our analysis. Additionally, Table 4-1 displays the most frequently occurring terms. Several of these terms can be frequently found in domains outside of nuclear activity.

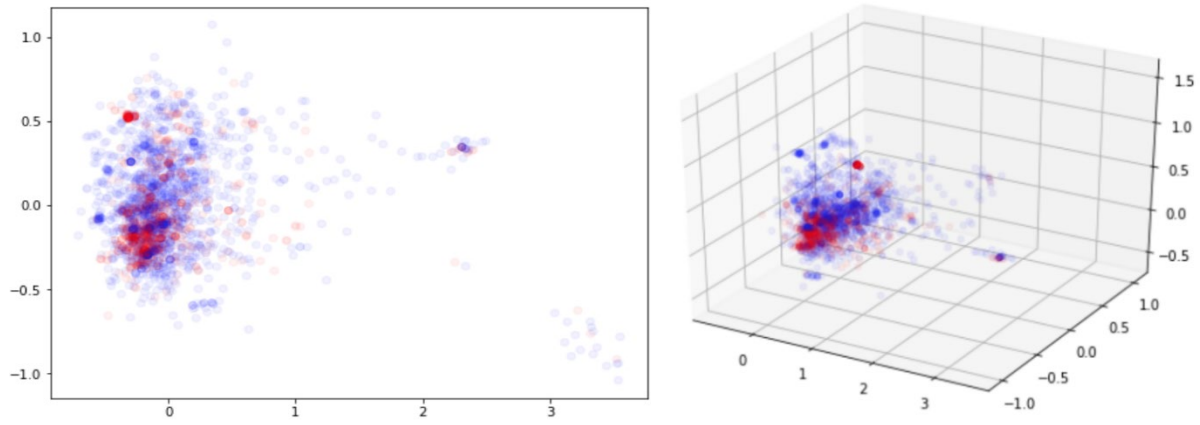

Figure 4-2: 2D and 3D Depictions of Word Embedding Models Created From the Initial Query Using the Full Glossary of Terms

Table 4-1: List of Most Frequent Terms From the Initial Query Using the Full Glossary of Terms

Term	Frequency	Term	Frequency
public_health	281008	air_conditioning	30945
federal_government	127432	annual_report	22105
national_security	99658	homeland_security	21796
executive_order	98081	cold_war	21504
parking_lot	59796	electrical_power	19875
local_government	52168	defense_secretary	19495
natural_gas	46182	office_space	14504
drinking_water	39798	risk_assessment	14434
unemployment_rate	39666	water_supply	14145
information_technology	33023	mortality_rate	13962

Based on this initial analysis the researchers chose to utilize a query that required documents identified using the most frequent terms to also contain the term "nuclear." The finalized query returned greater than 12 million articles. The time period covered by the query is January 2012 through May 2018. The resulting distribution of this query by day is shown in Figure 4-3. The quantity of articles in the entire Webhose database prior to 2015 is significantly smaller which resulted in fewer queried articles for that time frame.

The vast majority of articles are returned in the English language and therefore, no language filtration was deemed necessary.

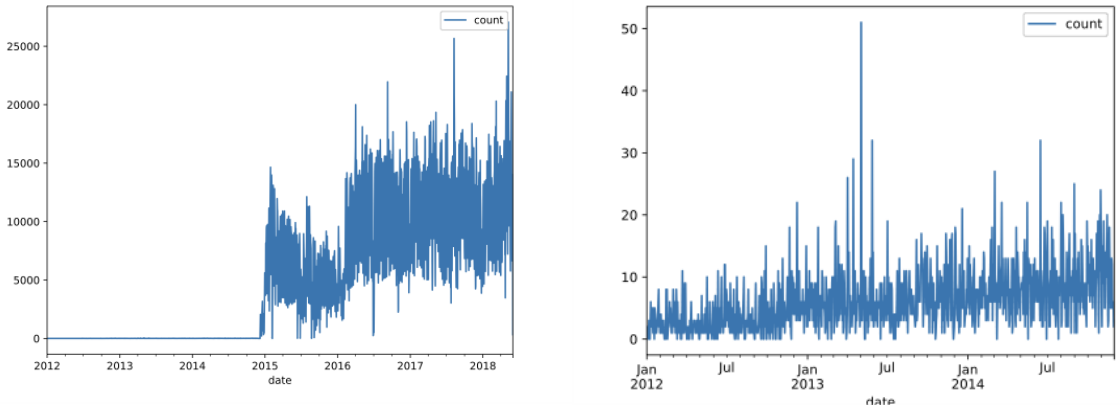

Figure 4-3: (Left) the Quantity of Query Results for the Entire Time Frame and (Right) the Quantity of Query Results Magnified for the Timeframe Prior to 2015

4.2 Word Embedding

4.2.1 Model Description

Given a set of unprocessed news articles *S* and a set of seed-phrases of interest *P*, provided by a domain expert, articles are preprocess to remove stop-words and also perform lemmatization and tokenization. This preprocessed set of documents (i.e., news articles) is subsequently used for further analysis. The next step is to transform the set of preprocessed documents *D* and phrases *P* (similarly preprocessed) into a common space allowing for similarity / distance calculations between articles and phrases. The popular technique of word embedding, specifically word2vec, is used to accomplish this task.

Word2vec models that are trained using a skip-gram model are used. Given a specific word from the text of a document (a news article in our case), the model tries to predict the words around (i.e., in the context of) the input word. This approach (popularly referred to as self-supervised learning) does not require the creation of any additional labelled data while allowing the effective learning of latent representations (i.e., a vector of real numbers) for each word in each news article in the corpus of documents *D*. The latent representation of words with similar meanings (or words related to each other) are similar, i.e., words that are closer in meaning are embedded closer together in the latent space than words that are farther in meaning.

The result of the word-embedding process yields an embedding-vector (*ev*) for each word in *D* and *P*. Using these *ev*'s, the overall embedding per document or phrase is derived as the mean *ev* of all the word embedding vectors therein. Thus, enabling the transformation of preprocessed news article data *D* and expert provided seed-phrases of interest *P* into a common latent space. We use a 100-dimensional latent *ev*'s for our experiments.

4.2.2 Evaluation

Word-embedding *ev*'s (also known as distributed representations) are known to intuitively represent the "meaning" of a word which is distributed across the entire *ev*. These *ev*'s are known to capture non-trivial and complex relationships between words, phrases and documents.

In Figure 4-4, the *ev*'s are visualized for each seed-phrase to characterize the representation learned by the word2vec embedding pipeline. The embeddings for each phrase are reduced from 100 to 2 dimensions using a dimensionality reduction technique like principal-component analysis. In the figure, four distinct groupings of phrases are highlighted, each indicated by a different colored box.

Phrases like {*nonproliferation treaty, nuclear weapon complex, stockpile need*} are all grouped close together into a latent topic we may deem to be related to *nuclear weapon proliferation*. Phrases {*regulatory requirement*, *requirement doe*, *federal regulation*, *federal program*} indicated by the red box all may be characterized by the topic *federal government agencies and programs related to nuclear science*. Similarly, the topics in the purple and the green boxes may be seen as corresponding to *nuclear proliferation logistics* and to *national-security and defense* respectively.

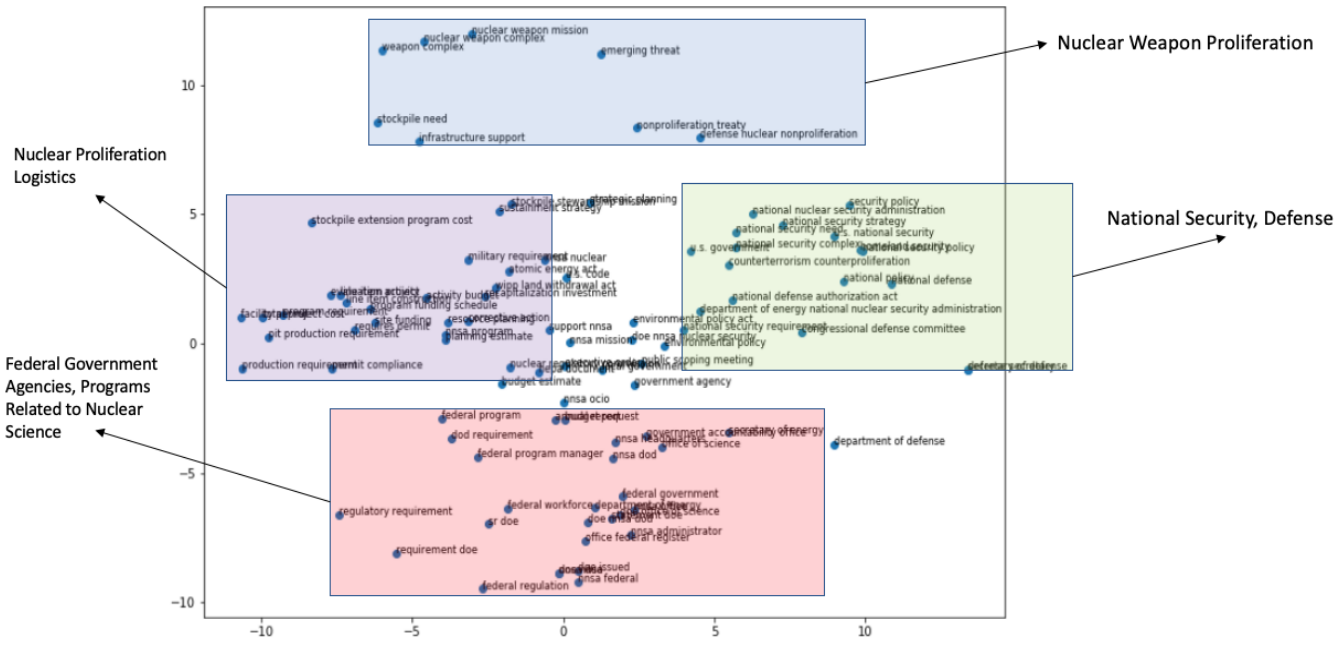

Figure 4-4: Word Embedding Seed-Phrases

In addition to grouping similar phrases, due to the common embedding space for words, documents and phrases, operations to retrieve the most similar words (along with the degree of similarity) were used for a given word or phrase. Figure 4-5 shows the ability of the word-embedding model to retrieve the most similar words given a query word or phrase. The results indicate that in each case i.e., for phrases of differing length, the model is able to retrieve different and relevant results indicating that it is not significantly affected by phrase length.

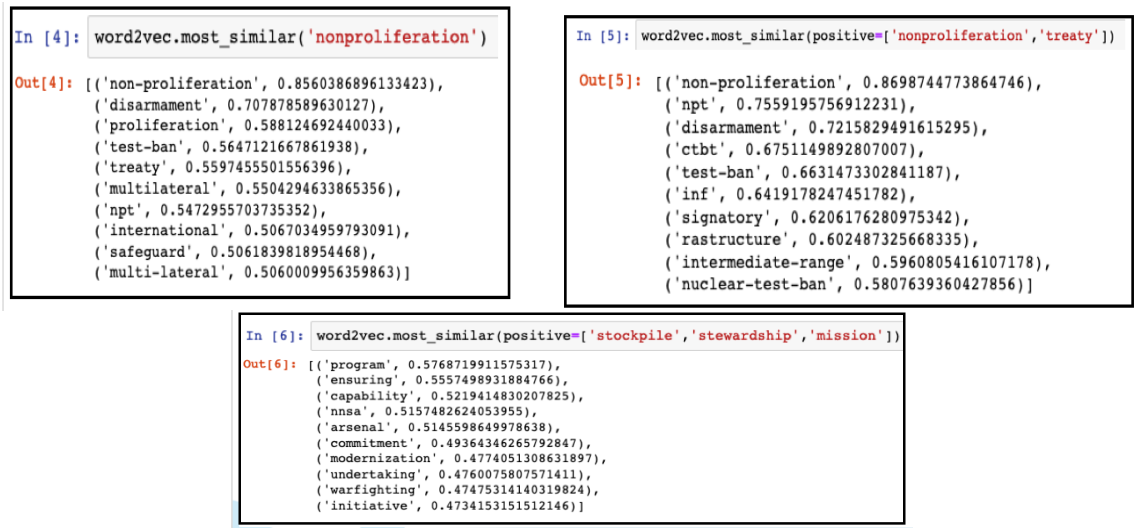

Figure 4-5: Word Embedding Top-K Most Similar Words

The learnt embeddings also yield non-trivial representations of the input query which is demonstrated in the example in Figure 4-6 where we see that the top words most similar to the phrase pit-production are not a mere superset of the most similar words for pit and production queried separately. This indicates that the word embeddings learn non-trivial representations based on the meaning and context of usage of each word and hence are able to effectively capture the topics represented in each phrase and document.

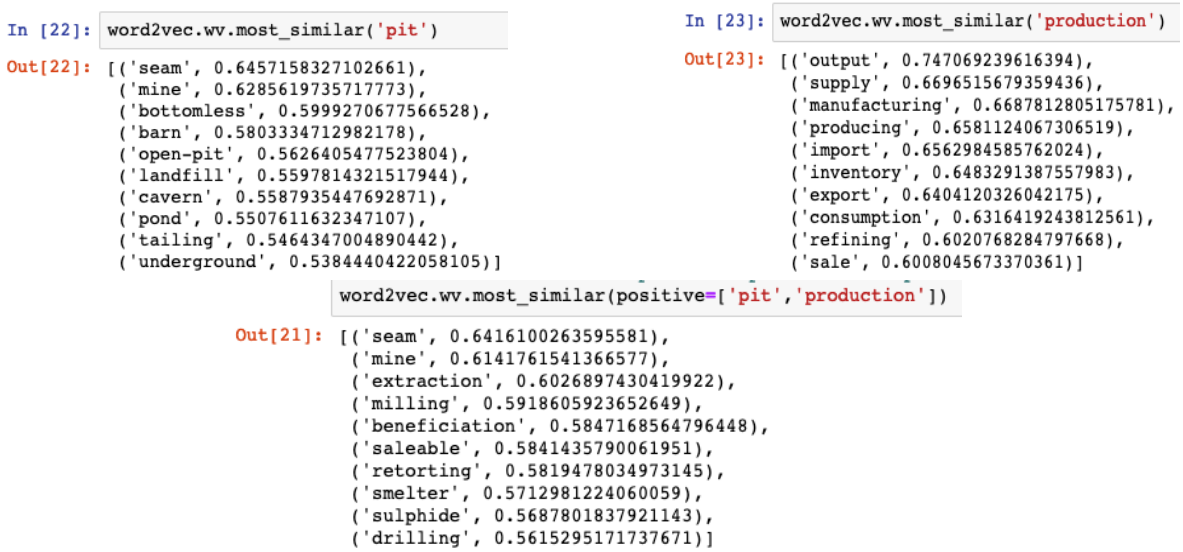

Figure 4-6: Word Embeddings Capture Meaningful-Relationships for Words & Phrases

Thus far, the research has characterized the phrase representations and showcased how seed-phrases are grouped together. In Figure 4-7, the researchers characterize a reduced dimensionality representation of the five most similar words (i.e., from set *D*) to each seed-phrase. In the figure each seed phrase point has a red border. The three seed phrases are *nonproliferation treaty*, *counterterrorism counterproliferation*, and *nnsa administrator*. It is apparent that the top-k most relevant words for each phrase uncover many interesting patterns. For example: the phrase *nonproliferation treaty* has as a close neighbor the word *disarmament* which is essentially synonymous with *nonproliferation*. Another neighbor is the acronym *np*

which is the acronym for our phrase of interest. Another property of interest is that the word embeddings also retrieve other related treaties like *ctbt* which is the *Comprehensive Nuclear-Test-Ban Treaty* which is essentially a treaty banning all nuclear explosive tests (i.e., generally relevant to nonproliferation but not specifically SRPPF). Similarly, notice that *sandia* (SNL) which is an NNSA affiliated laboratory is embedded close to the phrase *nnsa administrator*. Other subtle relationships like various sister programs related to and part of the NNSA like the advanced simulation and computing (*asc*) program are also embedded close to the phrase *nnsa administrator*.

This confirms that the word embeddings capture very sophisticated relationships between words and thereby capture a rich representation of the text in each document and phrase.

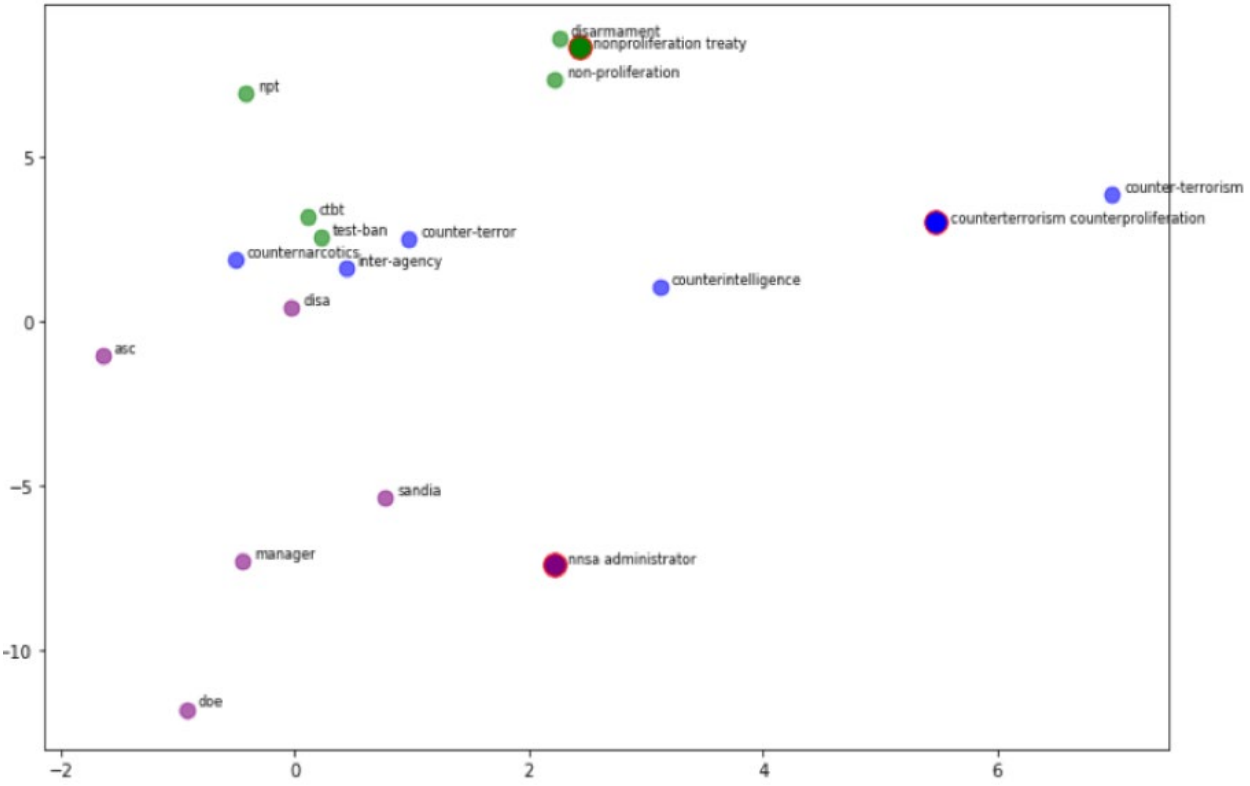

Figure 4-7: Top-N Similar Words Per Seed Phrase

4.3 Ranking

4.3.1 Model Description

Two ranking models, Embedding Ranking and BM25 Ranking, are described in the following sections. Following the individual descriptions, a fused model approach is described.

4.3.1.1 Embedding Ranking

The word-embedding pipeline obtains vector representations for documents and phrases in a common latent space. This latent representation for each document and phrase calculates similarities of every document, phrase pair, thereby inferring the overall similarity of a document to the set of seed phrases P. This overall score per document is used to order documents by similarity effectively ranking articles by their degree of similarity (i.e., relevance) to the topics of interest (given by P).

4.3.1.2 BM25 Ranking

While the embedding-based ranking approach captures semantic representations of documents to phrases, the phrases in P, being short snippets of text may lack enough semantic information for the embedding based method to perform effectively when used individually. Hence, the researchers augment the word-embedding based ranking algorithm developed in the previous step with a keyword based ranking algorithm by employing the BM25 algorithm as described in (Robertson and Zaragoza 2009). The algorithm operates as follows:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

Here, D represents a document and Q represents a query (i.e., a phrase in our case) consisting of one or more words. q_i represents the i^{th} word of query Q. For each q_i in Q, the inverse-document frequency IDF(q_i) is calculated. The IDF of a word is the inverse of the number of documents the word appears in, in each corpus. Intuitively, the IDF can be considered a penalty against frequently used words. For the research presented in this report, the IDF of each query word q_i is calculated using the training corpus of news articles (although we may also use pre-computed IDFs calculated using other text corpora). The function $f(q_i, D)$ returns the frequency of the query term in document D. $|D|$ represents the total length of the document while k_1 and b are hyperparameters. Thereby the BM25 algorithm yields a phrase document similarity score by summation of weighted frequency scores for each word (q_i), document (D) pair given a query Q. Biases due to document size and word frequency differences are explicitly accounted for by $|D|$ and IDF(-) calculations.

4.3.1.3 Fused Ranking

The BM25 ranking algorithm has a complexity $O(|Q|*|D|)$ where $|Q|$ indicates the phrase length and $|D|$ the document length. The algorithm can prove expensive for evaluation on the entire text corpus. Hence, the researchers first use the relatively inexpensive embedding-based rankings to obtain a basket of top relevant articles R. The BM25 ranking is then applied on this basket R_{emb}. The scores obtained from the BM25 and the embedding based rankings per document are then aggregated by an aggregation function agg() to obtain a Fused Ranking Score (FRS). The agg() function is an unweighted linear combination of the embedding based score and the BM25 score but more sophisticated functions may be used to fuse the ranking scores. The set of documents are then ranked by their FRS to obtain the final ranked list of articles R_{fused}. A sample of phrases and the corresponding results i.e., top 10 articles deemed most similar to each phrase follows below.

```
==========================================================================
Phrase = department of energy
DOE's Grid Study Clinging to the Past
Feds OK Georgia Power management takeover at plant Westinghouse is building
Oakridge’s Department of Energy East Tennessee Technology Park Recognized for Saving Resources and Taxpayer Money
ttps://t.co/9rRXNoF9y6
Feds OK Georgia Power management takeover at Plant Vogtle
Eck Industries exclusively licenses cerium-aluminum #alloy co-developed by ORNL
Over one thousand new microbial genomes discovered
Grid study Natural gas the leading cause for decline of coal not renewables
DOE approves Citicore's five hydropower service contracts totaling 2300 megawatts
Philippines' DOE gives go ahead for Citicore to pursue five hydropower projects
US officials aware of possible hacking at nuclear facilities
```

```
==========================================================================
Phrase = nnsa dod
HASC Subcommitee's Fiscal 2018 NDAA Bill Proposes Space-Focused Military Service
Second round of B61-12 qualification flight tests
Cost and safety at LANL
RadResponder Network Applications
Federally Funded Research And Development Center
World War 3? US tests 'most dangerous nuclear weapon ever produced' amid North Korea row
Blunders at LLNL
Luján: Safety Of LANL Workers And Los Alamos Community Must Be Paramount
United Launch Alliance wins launch of Air Force STP-3 mission
Intelligence Research Specialist job in Washington, D.C.
==========================================================================
Phrase = federal workforce
White House Wants to Increase Federal Employee Buyout Payments
National Technologies Associates, Inc. (NTA) Becomes V3 Virginia Values Veterans Program Certified
Colstrip gets $4.6 million in federal aid to retrain workers
Four States to Build Out Rural Broadband Infrastructure with Federal Grant Money
Job service offices to close in Anaconda, Dillon, Hamilton and Lewistown
H.R. 338, a bill to promote a 21st century energy and manufacturing workforce
Colstrip gets $4.6 million in federal aid to retrain workers - Tue, 01 Aug 2017 PST
Trump's Responsible Decision to End an After-School Program That Harms Children
Funds Boost Southeast's Entrepreneurial, Business Training
(USA-TX-HOUSTON) Clinic Office Manager
==========================================================================
Phrase = national security complex
Big Bite: Department of Defense Key Mission Challenges
FG Eulogised Nigerian Navy For Patronising Local Manufacturers
(USA-CO-Littleton) Mechanical Engineer
(USA-CO-Boulder) Systems Engineer Sr Stf
Allies, Partners Observe Cyber Guard Exercise
Aid and Stabilization in Afghanistan: What Do the Data Say?
(USA-VA-Manassas) Manassas, VA- Systems Engineer Intern
AnaVation Wins RS3 Contract with United States Army
(USA-VA-Manassas) Manassas, VA - Systems Engineer Associate
Oracle EBS Developer
==========================================================================
```

In the article list for the phrase national security complex, it can be seen that many job postings have been retrieved. One of the important indicators of nuclear proliferation is the hiring of a focused set of experts and hence a ranking model that is able to successfully identify and retrieve relevant articles of multiple types, even job postings is preferable. Overall, the top ranked articles per-phrase are relevant to the phrase. The per-phrase scores are aggregated to obtain the total fused similarity score of the article for the final overall ranking R_{fused}.

In addition to qualitative characterization of the ranking pipeline, an experiment is conducted soliciting expert ratings of the document quality. The experiment conducted was geared towards quantifying the degree of relevance of the ranking results produced by:

(i) The embedding-based ranking pipeline.
(ii) The BM25 + embedding-based ranking pipeline (i.e., fused ranking pipeline).

4.3.2 Evaluation

Two experts were each asked to rate the top-150 ranked articles in each of the two pipelines described in the previous section with ratings from 1 – 5 where a score of 1 indicates that the article is irrelevant and 5 indicates a high degree of relevance. To minimize bias, the experts were kept unaware of which pipeline generated each set of ranking results and the articles were also presented out of order for rating. The rated results were accumulated and used to calculate ranking metrics. We employed two popular ranking metrics to evaluate the quality of our ranking algorithm namely (Järvelin and Kekäläinen 2002) (Croft, Metzler and Strohman 2010):

(i) Discounted Cumulative Gain (DCG)

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)}$$

(ii) Normalized-Discounted Cumulative Gain (nDCG)

$$\mathrm{nDCG_p} = \frac{DCG_p}{IDCG_p}$$

The DCGp indicates the discounted cumulative gain obtained from ranking *p* documents. For each raked document, the relevance of the i^{th} document is represented by rel_i and is denoted by a score (i.e. 1 – 5). The nDCGp is the normalized version of the discounted-cumulative gain where the denominator indicates the Idealized-DCGp. The IDCGp is calculated by sorting the documents by the reviewer ratings to represent the idealized set of rankings. The Table 4-2 showcases the comparative DCG and nDCG values for 150 articles ranked using the embedding-based ranking and separately by the fused ranking scheme (i.e., BM25 + embedding). The fused ranking method significantly outperforms the embedding-based method in both the DCG and nDCG contexts indicating better ranking quality.

Table 4-2: Comparison of Model Ranking Quality for 150 Articles

Metric \ Model	Embedding-based Ranking	Fused Ranking
DCG150	56.46	62.81
nDCG150	0.8774	0.8934

The researchers further analyzed the top-k ranking quality from k = {1,2, ..., 150} and Figure 4-8 shows the nDCG value plotted for each case. The green curve indicates the fused ranking model while the blue curve indicates the embedding-based ranking model nDCG. We notice that the fused ranking model is significantly better than the embedding based ranking model throughout the course of the top-k ranking process.

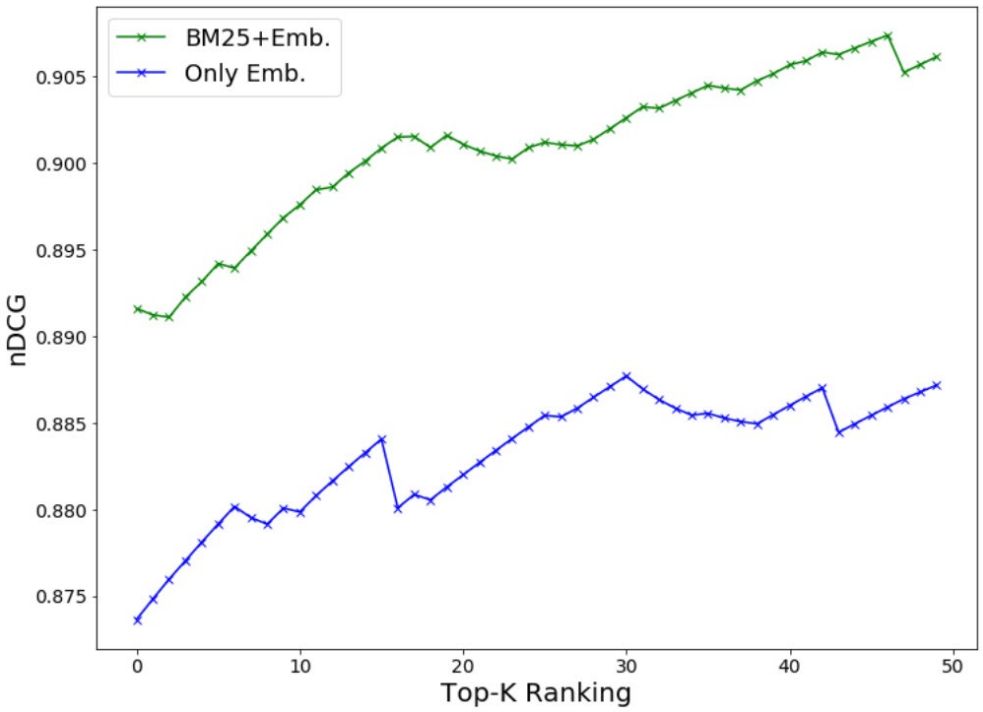

Figure 4-8: The nDCG value for each case

These results showcase the overall effectiveness and superiority of the fused ranking pipeline and therefore only rankings obtained from this pipeline are used moving forward.

4.4 Entity Extraction

4.4.1 Model Description

Thus far this report has discussed a pipeline to capture latent representations of documents and phrases using word embeddings. However, the objective also requires that the research identify overt themes discussed in each document to explicitly track certain topics of interest. The set of top ranked documents $DR << |D|$ are enriched by developing a named entity recognition module. A pre-trained model is employed from the popular natural language processing library SpaCy for this purpose. Named entity recognition (NER) is a sub-task of information extraction which essentially seeks to locate and classify "named entities" into pre-defined specific categories like Person, Location, Time, etc. This application focuses on the categories shown in Table 4-3.

Table 4-3: Entity Categories

Entity Category	Description
NORP	Nationalities or Religious or Political Groups
FAC	Buildings, Airports, Highways, Bridges
ORG	Companies, Agencies, Institutions
GPE	Countries, Cities, States etc.
PRODUCT	Objects, vehicles, foods, etc. (Not services)

The NER model is used to enrich all documents *DR*. Figure 4-9 shows an example document enriched with NER. Notice that the NER model employed is effectively able to identify GPE (geo-political entities), and persons mentioned in the document. It also effectively identifies other categories (not of specific interest to this research) like CARDINAL which essentially indicate mentions of quantities in the article. Also, of future potential interest would be the DATE entity although this is not employed currently.

4.4.2 Evaluation

The entity extraction pipeline is characterized in the next section with qualitative analysis of entity co-occurrence and by highlighting a few interesting examples of entity co-occurrence.

Figure 4-9: An Example Document Enriched With NER

4.5 Anomaly Detection – Entity Characterization

The following sections discuss consolidation of the entity and embedding enriched news articles into a searchable graph database.

4.5.1 Model Description

4.5.1.1 Initial Entity Characterization

In line with the first of two goals (i.e., to characterize the current state and evolution of entities of interest), a graph representation of our entities is developed by linking co-occurring entities (i.e., entities occurring in the same document) with a weighted edge. The weight of each edge is governed by the number of times a pair of entities co-occur. This weighted entity graph G may be constructed such that each graph G^t is constructed considering articles published only from time `$t-w$` to `$t$` where `$w$` may be a user-governed time window of consideration. Figure 4-10 and Figure 4-11 represent various sub-graphs of one such G^t constructed using entities from articles published between Jun – Aug 2017. Figure 4-10 showcases the ego-network (top 50 co-occurring entities) of a query of interest (i.e., *nrc*) representing the Nuclear Regulatory Commission. This ego-network based entity representation allows us to focus explicitly on specific entities of interest and yield a multi-faceted view in terms of the `GPE`, `FAC`, `NORP` and `ORG` that are in the neighborhood (and hence highly related) to the entity of interest at the selected time slot. In Figure 4-10, notice there are entities of multiple types that yield a heterogeneous entity co-occurrence graph. This view allows an analyst to glean information about specific organizations, geopolitical entities or specific products etc. most relevant to the entity of interest in the current time period.

Figure 4-11 characterizes the ego-network for another entity *sandia*. The ego-network shows many national laboratories mentioned as co-occurring entities which is intuitively acceptable. Another interesting occurrence is the mention of the entities *heliobiosys inc.*, *cyanobacteria*, and *marine cyanobacteria* where these entities highlight:

(i) a collaboration between Sandia national laboratories and Heliobiosys Inc., and
(ii) the involvement of Sandia national laboratories in projects based on cyanobacteria.

Upon further investigation for news articles involving Heliobiosys Inc., (a company extensively dealing with cyanobacteria) and Sandia National Laboratories, a news article was identified that highlights the collaboration between Sandia National Laboratory, Heliobiosys Inc. and Lawrence Berkeley National Laboratory (commonly referred to as Berkley Lab) on a project employing cyanobacteria for biofuel production. This highlights that the heterogeneous entity ego-network highlights surprising insights and developments involving entities of interest.

These ego-networks serve as a qualitative validation of the entity extraction model. The graphs display expected and reasonable connections.

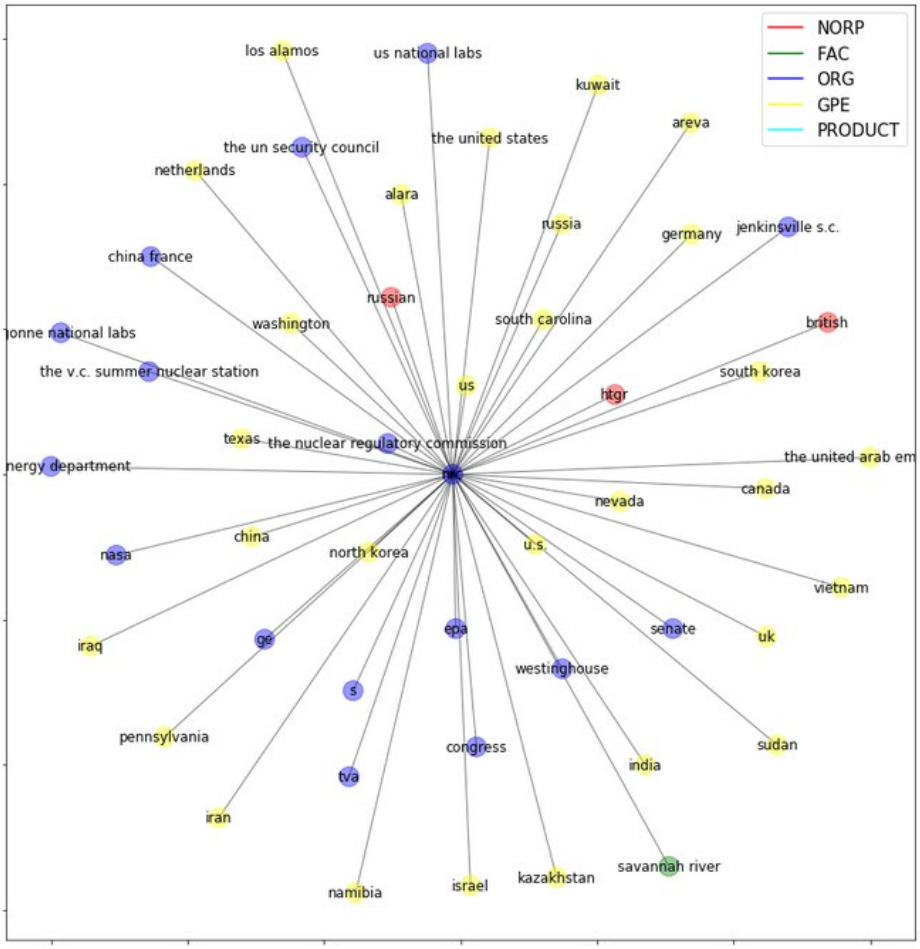

Figure 4-10: Heterogeneous Entity Ego-Network (*nrc*)
The distances and positions of nodes are arbitrary and have been chosen only for visual clarity.

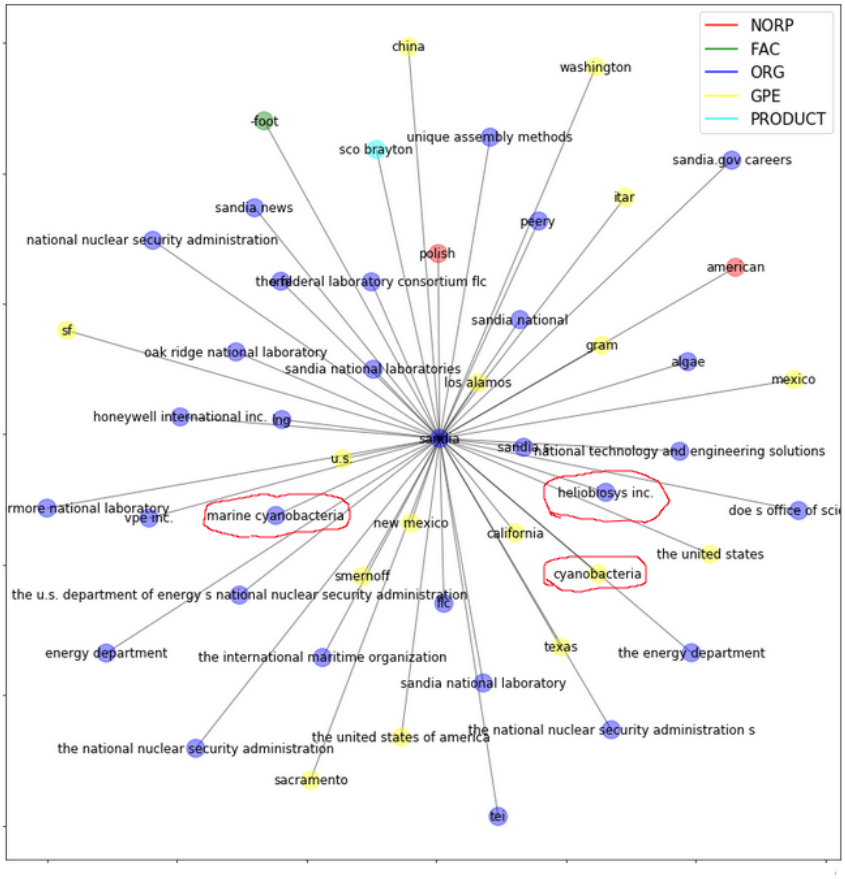

Figure 4-11: Heterogeneous Entity Ego-Network (*sandia*)
The distances and positions of nodes are arbitrary and have been chosen only for visual clarity.

4.5.1.2 Entity Characterization – Temporal Evolution (Continuing Development)

The above section described static entity ego-networks. Another aspect of entity characterization involves understanding the state of change of the entity (in terms of its ego-network) over time. This may be conducted by calculating the degree of change in an entity's state over successive time periods. For example, consider an entity's state to be represented by a multi-hot encoded vector corresponding to the IDs of other entities in its ego network. A notion of change may be obtained by comparing the Jaccard similarity (i.e., similarity of binary multi-hot entity state vectors) over successive time periods. Such a characterization performed for all entities of interest would allow the population of the most "chaotic" entities (i.e., entities that exhibit the greatest state change).

4.5.1.3 Anomaly Detection (Continuing Development):

Finally, the researchers will also address the anomaly detection task by leveraging the current entity extraction pipeline and entity co-occurrence network. The anomaly detection problem will be treated as one of link prediction wherein it predicts edges occurring between entities for a particular entity of interest and unexpected edges or non-existent expected edges may be deemed anomalous occurrences. The temporal evolution characterization and anomaly detection pipeline are under development.

5.0 Conclusions

Models and methods have been developed to detect organization names, facility, and function identification as well as events within the openly available data.

Analysis of Twitter™ data demonstrates that word embedding models trained on data obtained from queries using the glossary of terms produce an interpretable signal of key terms and the connectedness to relevant entities that automated models can use. The research team is still evaluating current methods to determine the best method for use in the prototype system. For example, a time dependent word embedding algorithm that retains all cumulative data may be advantageous in some cases that creates a strong link to an event that occurs earlier or later in time. However, keeping all data could create too low of a "signal-to-noise" ratio and prohibit straightforward detection of changes of interest. A "rolling window" can help to eliminate noisy data by shifting over time, however, the window size or rolling length needs tuning to avoid missing key events that are linked across time. Next, the research team plans to use this analysis to build a prototype system that can identify the indicators of nuclear activity.

For the news article data, elements of an automated preliminary data pipeline have been built and the evaluation of each step was presented above. After word embedding, ranking and entity extraction strong relationships and indication of entity detection were shown. An anomaly detection model that analyzes the temporal evolution of the resulting networks is still under development. The identified anomalies will be the raw output of the model which will feed into a fusion model.

Each algorithm has presumed advantages and disadvantages, as well as parameters that need to be tuned and characterized for application in the prototype model development. The second year of the project development will develop fusion models to leverage the advantages of the individual models. A prototype system will be created and demonstrated in the next year of the project.

6.0 References

Ahmad, Subutai, Alexander Lavin, Scott Purdy, and Zuha Agha. 2017. "Unsupervised real-time anomaly detection for streaming data." *Neurocomputing* (262): 134-147. doi:10.1016/j.neucom.2017.04.070.

Arterburn, Jason, Erin D. Dumbacher, and Page O. Stoutland. 2021. *Signals in the Noise: Preventing Nuclear Proliferation with Machine Learning & Publicly Available Information.* Washington, D. C., USA: Nuclear Threat Initiative. Accessed January 19, 2021. http://nti.org/289R.

Croft, W. Bruce, Donald Metzler, and Trevor Strohman. 2010. *Search engines: Information retrieval in practice.* Vol. 520. Addison-Wesley.

Danielson, T. L., A. A. Kail, and J. A. Pike. April 2020. *Event Definition for the Automated Detection of Nuclear Proliferation Activities.* Aiken, SC: Savannah River Natinal Laboratory, SRNL-STI-2020-00155, Rev. 0.

DOD. 2020. *Nuclear Matters Handbook 2020.* The Office of the Deputy Assistant Secretary of Defense for Nuclear Matters.

DOE. April 2020. *Draft Environmental Impact Statement for Plutonium Pit Production at the Savannah River Site in South Carolina.* Savannah River Site: U. S. Department of Energy National Nuclear Security Administration, DOE/EIS-0541.

DOE. May 2003. *Draft Supplemental Programmatic Environmental Impact Statement on Stockpile Stewardship and Management for a Modern Pit Facility.* U. S. Department of Energy National Nuclear Security Administration, DOE/EIS-236-S2.

DOE. April 2014. *Fiscal Year 2015 Stockpile Stewardship and Management Plan Report to Congress.* Washington, D.C.: United States Department of Energy.

DOE. March 2015. *Fiscal Year 2016 Stockpile Stewardship and Management Plan Report to Congress.* Washington, D.C.: National Nuclear Security Administration United States Department of Energy.

DOE. March 2016. *Fiscal Year 2017 Stockpile Stewardship and Management Plan – Biennial Plan Summary Report to Congress.* Washington, D.C.: National Nuclear Security Administration United States Department of Energy.

DOE. November 2017. *Fiscal Year 2018 Stockpile Stewardship and Management Plan Report to Congress.* Washington, D.C.: National Nuclear Security Administration United States Department of Energy.

DOE. October 2018. *Fiscal Year 2019 Stockpile Stewardship and Management Plan – Biennial Plan Summary Report to Congress.* Washington, D.C.: National Nuclear Security Administration United States Department of Energy.

DOE. July 2019. *Fiscal Year 2020 Stockpile Stewardship and Management Plan Report to Congress.* Washington, D.C.: National Nuclear Security Administration United States Department of Energy.

—. n.d. *Maintaining the Stockpile.* Accessed January 19, 2021. www.energy.gov/nnsa/missions/maintaining-stockpile.

—. n.d. *Plutonium Pit Production National Nuclear Security Administration.* Accessed January 19, 2021. www.energy.gov/nnsa/plutonium-pit-production.

Goldstone, Jack A., Robert H. Bates, David L. Epstein, Ted R. Gurr, Michael B. Lustik, Monty G. Marshall, Jay Ulfelder, and Mark Woodward. 2010. "A Global Model for Forecasting Political

Instability." *American Journal of Political Science* 54 (1): 190-208. doi:10.1111/j.1540-5907.2009.00426.x.

Hu, Mengdie, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. 2012. "Breaking News on Twitter." *SIGCHI Conference on Human Factors in Computing Systems.* New York, NY, USA: ACM. 2751–54. doi:10.1145/2207676.2208672.

Järvelin, Kalervo, and Jaana Kekäläinen. 2002. "Cumulated gain-based evaluation of IR techniques." *ACM Transactions on Information Systems (TOIS)* (ACM) 20 (4): 422-446.

2018. *Joint Statement from Ellen M. Lord and Lisa E. Gordon-Hagerty on Recapitalization of Plutonium Pit Production.* May 10. https://www.energy.gov/nnsa/articles/joint-statement-ellen-m-lord-and-lisa-e-gordon-hagerty-recapitalization-plutonium-pit .

Lunetta, Ross S., Joseph F. Knight, Jayantha Ediriwickrema, John G. Lyon, and L. Dorsey Worthy. 2006. "Land-Cover Change Detection Using Multi-Temporal MODIS NDVI Data." *Remote Sensing of Environment* 105 (2): 142–54. doi:10.1016/j.rse.2006.06.018.

O'Brien, Sean P. 2010. "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research." *International Studies Review* 12 (1): 87–104. doi:10.1111/j.1468-2486.2009.00914.x.

Olson, Michael, Annie Liu, Matthew Faulkner, and K. Mani Chandy. 2011. "Rapid Detection of Rare Geospatial Events: Earthquake Warning Applications." *Proceedings of the 5th ACM International Conference on Distributed Event-Based System.* New York, NY, USA. 89–100. doi:10.1145/2002259.2002276.

Ramakrishnan, Naren, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, and et_al. 2014. "'Beating the News' with EMBERS: Forecasting Civil Unrest Using Open Source Indicators." *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM. 1799–1808. arxiv.org/abs/1402.7035.

Ranshous, Stephen, Shitian Shen, Danai Koutra, Steve Harenberg, Christos Faloutsos, and Nagiza F. Samatova. 2015. "Anomaly Detection in dynamic networks: a survey." *WIREs Computational Statistics* 7: 223-247. doi:10.1002/wics.1347.

Rekatsinas, Theodoros, Saurav Ghosh, Sumiko R. Mekaru, Elaine O. Nsoesie, John S. Brownstein, Lise Getoor, and Naren Ramakrishnan. 2015. "SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources." *Proceedings of the 2015 SIAM International Conference on Data Mining.* SIAM. 379–87. doi:10.1137/1.9781611974010.43.

Robertson, S, and H. Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond.* Now Publishers Inc.

Zhao, Liang, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. 2014. "Unsupervised Spatial Event Detection in Targeted Domains with Applications to Civil Unrest Modeling." *PloS One* 9 (10): e110206. doi:10.1371/journal.pone.0110206.

Appendix: Search and Glossary Terms

This appendix contains the list of terms used in the first step of relevant item identification and the specific event related glossaries.

Generic Search Terms Related to Pit Production

nuclear security	pit production capacity	tritium enterprise	material control
stockpile stewardship	national nuclear security	doe facility	pantex operation
national security	nuclear testing	lanl plutonium	proposed srppf complex
pit manufacturing capacity	u.s. nuclear weapon	design engineering	weapon assessment
national nuclear security administration stockpile stewardship	nuclear criticality	federal facility	sustainment program
pit nuclear weapon	war reserve	pit pantex	extension program lep
long-term pit	nnsa site	project proposal	maintain stockpile
weapon production site	pantex site	plutonium aging	feasibility study
nuclear security enterprise	nevada test site	proposed action alternative	independent cost estimate
draft environmental impact statement	nuclear stockpile	weapon modernization	scoping meeting
los alamos	support mission	manhattan project	pit aging
savannah river	oak ridge	weapon mission	carlsbad new mexico
life extension	pantex plant	weapon dismantlement disposition	aiken county
weapon stockpile	stockpile responsiveness	weapon reliability	warhead life extension
department of energy stockpile stewardship	directed stockpile work	Savannah River Site boundary	nuclear weapon performance
pit production environmental impact statement	support stockpile	weapon life cycle	stockpile maintenance
sr pit production	srppf complex	plutonium sustainment	support pit production
nuclear weapon stockpile	stockpile assessment	stockpile sustainment	global security
carlsbad site	current stockpile	proposed alternative	tennessee valley authority
weapon activity	sandia national laboratory	lanl operation	pit production process
weapon system	stockpile weapon	stockpile surveillance	address aging
national security laboratory	lawrence livermore national laboratory	support stockpile stewardship	pit reuse
life extension program	nuclear nonproliferation	amarillo texas	active stockpile
nevada national security site	limited life component	rocky flats plant	pit production mission
stockpile stewardship program	weapon dismantlement	oak ridge tennessee	infrastructure nnsa
defense program	enduring stockpile	albuquerque new mexico	weapon design cost
pit manufacturing	cold war	required support mission	weapon life extension
doe order	waste isolation pilot plant	energy defense	stockpile warhead
nuclear posture	national archive record	stockpile modernization	disposition program
nuclear posture review	dismantlement disposition	pantex sr	weapon surety
production capacity	pit lifetime	warhead life	strategic partnership
savannah river site	los alamos site	program milestone	u.s. nuclear stockpile
production capability	safeguard security	future weapon	plutonium strategy

Generic Search Terms Related to Pit Production

nuclear security program	stockpile responsiveness program	mission support	stockpile support
proposed srppf	disposition plutonium	doe/nnsa nuclear	hanford site
los alamos national laboratory	nuclear weapon life cycle	barnwell county	savannah river plutonium
pit facility alternative	wipp site	weapon testing	material evaluation
nuclear weapon council	stockpile evaluation	stewardship mission	material aging
doe site	nuclear safety	nuclear counterterrorism	north augusta
production rate	stockpile stewardship u.s.	existing stockpile	modern nuclear weapon
deferred maintenance	site screening	program milestone objective	augusta georgia
federal agency	public scoping	modernization program	richmond county
weapon performance	capacity requirement	plutonium operation	
nuclear deterrent	production capability lanl	pit production lanl	

Glossary of Terms Related to Economic Events

The following list of terms fall under "Economic Events", which includes the sub-domains: "Stock Market", "Corporate Developments", and "Global Currency Trends".

Glossary of Terms Related to Economic Events

supporting industry	average earnings	employment income	current housing market
housing market	facility additional jobs	low unemployment	
new job created	created supporting industry	low unemployment rate	
indirect job created	generate additional indirect income	unemployment rate roi	

Glossary of Terms Related to Acquisition Events

The following list of terms fall under "Acquisition Events", which includes the sub-domains: "Non-Nuclear Material Acquisition", "Nuclear Material Acquisition", "Nuclear Technology Acquisition", and "Facility/Infrastructural Acquisition".

Glossary of Terms Related to Acquisition Events

nuclear weapon	highly enriched uranium	plutonium disposition	nondestructive evaluation
pit production	experimental capability	srppf operation	weapon simulation
pit facility	chemistry metallurgy	test equipment	uranium component
modern pit facility	pit production capability	waste treatment	material facility
tru waste	chemistry metallurgy research	sanitary wastewater generated	facility upgrade
hazardous waste	mixed waste	computing system	office space
radioactive material	physical infrastructure	lithium production	site construction
weapon component	simulation capability	access road	reactor fuel
nuclear material	pit manufacturing capability	disposal capacity	pit production facility
research and development	facility design	computational capability	llw disposal facility
construction activity	technology engineering	inactive warhead	shipment wipp
enriched uranium	high performance computing	aging infrastructure	offsite disposal
sensitivity analysis	engineering science	stockpile size	weapon assembly
plutonium pit	llw disposal	new pit	construction modification
construction operation	wastewater discharge	electrical energy	backup diesel generator
production facility	science technology engineering	test facility	nuclear weapon infrastructure
processing facility	storage facility	spent plutonium	responsive infrastructure
pit manufacturing	component production	plutonium fuel	aging facility
nuclear explosive	nuclear warhead	plutonium experiment	modeling capability
weapon production	feed preparation	infrastructure investment	advanced technology development
plutonium metal	component manufacturing	infrastructure modernization	infrastructure need
construction impact	programmatic infrastructure	major alteration	material receipt
new facility	hydrodynamic test	engineering capability	control facility
nuclear weapon production	molten plutonium	manufacturing development program	air conditioning
waste generated	fabrication facility	baseline cost report	pit disassembly
radioactive waste	waste stream	produce plutonium	ventilation system
sanitary wastewater	material characterization	storage vault	bulk storage
drinking water	storage tank	plant capacity	operational capability
advanced manufacturing	cooling tower	fuel fabrication facility	nnsa infrastructure
existing facility	fuel fabrication	proposed facility	pit produced
manufacturing capability	molten plutonium metal	radiological transportation	transportation route
facility operation	experimental facility	development activity	facility modification
waste generation	system component	reserve plutonium	parking lot
uranium processing	technology cybersecurity	war reserve plutonium	laboratory facility

Glossary of Terms Related to Acquisition Events

secure transportation	surplus plutonium	minor construction	weapon technology
mixed llw	material waste	acquisition report	mixed tru waste
weapon production facility	radiation protection	enrichment capability	construction cost
additive manufacturing	construction equipment	non nuclear component	enriched uranium material
hazardous chemical	analytical chemistry	facility constructed	scientific understanding
site infrastructure	electrical power	waste staging/tru	advanced technology system
hazardous material	weapon engineering	quantity plutonium	nuclear design
uranium processing facility	ballistic missile warhead	fuel processing	computing technology
technology development	cruise missile warhead	radiological material	lithium production capability
neutron generator	hepa filter	tru waste generated	waste processing
information technology	waste acceptance	diesel fuel	data analysis
solid waste	wastewater treatment	commercial facility	industrial effluent
science engineering	new building	radiological chemical	extraction facility
special nuclear material	manufacturing activity	doe modern pit	solid waste disposal
casting furnace	construction facility	spent plutonium fuel	modeling simulation capability
transportation asset	pit production rate	reserve plutonium pit	production capacity ppy
depleted uranium	equipment facility	production schedule	hvac exhaust stack
disposal facility	process development	enhanced capability	radioactive liquid waste
feed casting furnace	construction site	disposition facility	staging facility
weapon design	wastewater generated	weapon designer	portable toilet
future stockpile support facility	wrought process	fire protection	volatile organic compound
manufacturing development	computing initiative	design feature	wastewater treatment facility
construction project	information technology cybersecurity	nuclear weapon system	non -nuclear component
science program	simulation computing program	organic compound	science engineering capability
uranium enrichment	waste wipp	effluent discharge	surplus plutonium disposition
advanced simulation computing	manufacturing technology	nonhazardous waste	mixed oxide fuel
weapon program	water supply	waste shipment	produce plutonium pit
strategic material	natural uranium	waste disposed	pit manufacturing operation
explosion feed casting	plutonium pit manufacturing	weapon material	disposal tru waste
cyber security	capital acquisition	retired weapon	hydrodynamic testing
transfer system	shipping container	armored tractor	pyrochemical processing method
liquid waste	plutonium processing	advanced diagnostics	plutonium processing facility
treatment facility	llw generated	uranium material	facility modernization
material component	stormwater runoff	new manufacturing	knowledge transfer
tritium production	transported material	physics engineering	fissile component
nuclear facility	existing infrastructure	communication system	plutonium component
nuclear component	waste storage	technology production	conversion facility
physical security	transuranic waste	mixed oxide	plutonium material
federal workforce	material process	manufacturing facility	storage nuclear weapon
facility infrastructure	research reactor	exhaust stack	metallurgy research building
general purpose infrastructure	mission capability	shipment tru	device assembly facility

Glossary of Terms Related to Acquisition Events

fissile material	production activity	disposal site	material transportation
science technology	infrastructure requirement	operation building	material disposition
advanced technology	engineering controls	interim pit production	mixed oxide fuel fabrication
waste disposal	pit manufacturing activity	tru waste wipp	weapon-grade plutonium
facility equipment	processing method	waste staging/tru packaging	production infrastructure
nitrogen dioxide	security infrastructure	water system	refurbished warhead
material science	nnsa facility	site activity	uranium metal
gas transfer system	computing capability	prior construction	engineering facility
plutonium facility	system engineering	construction area	computing facility
waste facility	component subsystem	spent nuclear fuel	pit plutonium
engineering program	waste handling	nuclear weapon design	plutonium oxide
nuclear fuel	quality assurance	nuclear weapon program	plutonium americium
non-nuclear component	pyrochemical process	material science engineering	supporting infrastructure
domestic uranium	development testing	uranium enrichment capability	transportation radioactive material
domestic uranium enrichment	waste acceptance criterion	radiation -hardened	infrastructure recapitalization
protection equipment	wipp disposal	weapon component system	weapon component material
facility construction	hearing protection	analytical support	number pit produced
advanced manufacturing development	nuclear reactor	process equipment	disassembly conversion facility
diesel generator	plutonium pit production	construction new facility	analytical laboratory
natural gas	hydrodynamic experiment	new facility built	
metallurgy research	underground nuclear explosive	weapon infrastructure	
nuclear weapon component	engineering design	processing manufacturing	

Glossary of Terms Related to Political/Diplomatic Events

The following list of terms fall under "Political/Diplomatic Events", which includes the sub-domains: "International Diplomacy", "Domestic Diplomacy", and "Elections/Political Appointments".

Glossary of Terms Related to Political/Diplomatic Events

secretary of energy	regulatory requirement	national security need	stockpile stewardship mission
budget request	national defense authorization act	wipp land withdrawal act	total project cost
executive order	nnsa office	defense nuclear nonproliferation	recapitalization investment
secretary of defense	congressional defense committee	security policy	nuclear weapon mission
federal regulation	federal program	weapon complex	line item construction
federal workforce	planning estimate	statement doe	sr doe
nnsa dod	federal government	nepa document	strategic planning
national nuclear security administration	nuclear regulatory commission	requirement doe	evaluation activity
doe/nnsa	budget estimate	nnsa doe	national security policy
department of energy	stockpile need	resource planning	u.s. national security
u.s. government	nnsa nuclear	infrastructure support	nnsa administrator
department of energy/national nuclear security administration	nnsa headquarters	nnsa ocio	counterterrorism counterproliferation
office federal register	emerging threat	activity budget	defense secretary
u.s. code	government agency	program funding schedule	sustainment strategy
national defense	nnsa program	production requirement	site funding
nnsa mission	corrective action	program requirement	federal program manager
stockpile extension program cost	national security strategy	permit compliance	doe/nnsa dod
office of science	doe office of science	nnsa federal	line item project
environmental policy	dod requirement	requires permit	public scoping meeting
national security complex	atomic energy act	national policy	pit production requirement
environmental policy act	support nnsa	nuclear weapon complex	government accountability office
military requirement	doe issued	national security requirement	nonproliferation treaty
department of defense	doe nnsa	homeland security	facility permit
doe/nnsa nuclear security	local government	annual report	

Glossary of Terms Related to Population/Personnel Events

The following list of terms fall under "Population/Personnel Events", which includes the sub-domains: "Population Redistribution", "Accumulation of Skillsets", and "Environmental/Illness Events".

Glossary of Terms Related to Population/Personnel Events

environmental impact	occupational safety health	population living	risk mitigation
environmental impact statement	pollution prevention	program manager	nnsa sar
draft environmental impact statement	facility accident	latent cancer fatality	worker increase
pit production environmental impact statement	radiological release	total population	worker dose
air quality	injury illness	regional influence	site environmental report
health effect	risk reduction	construction industry	security enterprise workforce
air emission	radiological air emission	additional jobs created	regulatory limit
job created	public health	incident response	impact groundwater
fatal cancer	impact surface water	environmental restoration	nonradiological air emission
labor force	injury illness fatality	release radioactivity	environmental assessment
radiation exposure	radiological worker	lost workday	influx of new workers
air pollutant	protective force	nuclear radiation	new housing
water quality	radiation worker	population surrounding	housing demand
cancer risk	national environmental policy	emergency preparedness	chemical exposure
radiation dose	human health accident	potential risk	new housing demand
adverse impact	exposed individual	hazardous air pollutant	new resident expected
radiological impact	latent cancer	nonradiological impact	worker population
dose person-rem	fission event	potential environmental impact	contaminated soil
radiation dose equivalent	transportation accident	unemployment rate	labor force estimated
cancer fatality	environmental report	groundwater resource	acid release
ionizing radiation	dose offsite	drinking water standard	groundwater quality
accident scenario	dose rate	total workforce	mortality rate
radiological accident	federal employee	programmatic eis	environmental monitoring
fire-induced release	exposure guideline	exposure limit	hearing protection program
effective dose	natural resource	release radioactive material	human capital
chemical accident	environmental protection agency	traffic noise	annual radiation dose
material spill	chemical accident frequency	environmental effect	greenhouse gas emission
occupational safety	exposed offsite individual	risk assessment	radioactive spill
health impact	radiological accident frequency	occupational injury	site-wide environmental impact
air quality standard	concentration limit	thyroid cancer	workforce nnsa
radioactive material spill	material risk	workforce need	nuclear incident
environmental impact associated	potential accident	capability needed	construction employee
construction worker	chemical release	equipment failure	safety analysis
scientist engineer	injury fatality	emergency planning	design analysis
offsite population	radiation effect	pollution control	accident analysis

Distribution:

M. J. Barnes, mark.barnes@srnl.doe.gov
J. S. Bollinger, james02.bollinger@srnl.doe.gov
N. J. Bridges, Nicholas.Bridges@srnl.doe.gov
P. Butler, pabutler@vt.edu
G. R. Cefus, gregory.cefus@srnl.doe.gov
T. L. Danielson, Thomas.Danielson@srnl.doe.gov
C. C. Herman, connie.herman@srnl.doe.gov
C. M. Gregory clint.gregory@srnl.doe.gov
D. G. Jackson, Jr., dennis.jackson@srnl.doe.gov
R. B. James, Ralph.James@srnl.doe.gov
R. D. Jeffcoat, ron.jeffcoat@srnl.doe.gov
G. F. Kessinger, glen.kessinger@srnl.doe.gov
P. L. Lee, patricia.lee@srnl.doe.gov
B. Mayer, bmayer@cs.vt.edu
N. Muralidhar, nik90@vt.edu
P. R. Nuessle, patterson.nuessle@srnl.doe.gov
S. M. Oswald, sandi.oswald@srnl.doe.gov
J. A. Pike, jeff.pike@srnl.doe.gov
N. Self, nwself@vt.edu
T. Whiteside, Tad.Whiteside@srnl.doe.gov
D. L. Wilson, david.wilson@srnl.doe.gov
Records Administration (EDWS)