U.S. Department of Energy
Office of ES&H Reporting and Analysis: Similarity Search Use Cases and Applications

Presentation for the U.S. Nuclear Regulatory Commission
Data Science and Artificial Intelligence Regulatory Applications Workshops
August 18, 2021

Felix Gonzalez, P.E.
Office of ES&H Reporting and Analysis
U.S. Department of Energy
Felix.Gonzalez@hq.doe.gov
301-903-9311

Office of Environment, Health, Safety and Security

Presentation Agenda

- Overview of similarity search ranking process and Natural Language Processing (NLP)
- Applications and Use cases
  - Q&A's database
  - Query complex Environment, Safety and Health (ES&H) related topics
- Lessons Learned and Concluding remarks

Process Summary: Query and Similarity Ranking

Database (structured and unstructured data)

Data Filtering

Unstructured data (i.e., text) processing

Formulate query (e.g., input text, keywords, question, etc.)

Submit data filters and text query

Input text processing

BOW Model with TFIDF

Vectorization

Similarity Ranking (i.e., cosine)

Sorting Results

Ranked Results

User Interface

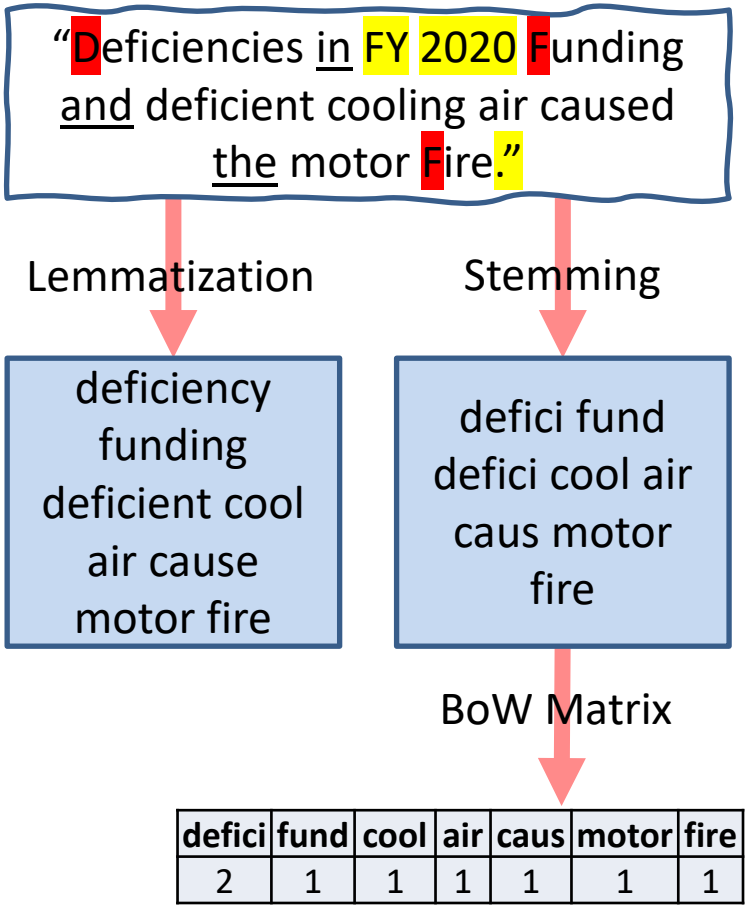
Natural Language/Text Processing

- Text processing and normalization:
  - Lower-case (red)
  - Removes special characters, numbers, 2-character words, etc. (Yellow)
  - Remove stop-word (underlined)
  - *Lemmatization or Stemming*

- Model and metrics used:
  - Bag of Words (BoW) model
  - Term Frequency-Inverse Document Frequency (TFIDF)

- BoW and TFIDF used to calculate the cosine similarity metric

Sample Text Normalization and BoW Matrix



defici	fund	cool	air	caus	motor	fire
2	1	1	1	1	1	1

Search Query Application: Q&A's database

- DOE's COVID Hotline has answered questions from staff since the start of the pandemic
  - Q&A'S were initially tracked via spreadsheet in a shared drive
  - Hotline representatives searched the spreadsheet for answers
- As the spreadsheet grew it became challenging to find answers to questions
- An application was developed to show potential of Chat Bots to support the Hotline operations
- Hotline representatives requested the application instead show the top results which would improve their efficiency in evaluating questions and obtaining an answer quickly
- Evolved into a similarity search application that was integrated into Hotline's existing framework

Q&A Database Example
Text Normalization (1/2)

Sample Q&A's in the database

Question	Answer
Will DOE be collecting personal information upon building re-entry and, if so, how will it be protected?	We are not currently collecting personal or health information, but if it is determined to become necessary, any personal and health information collected by DOE or its contractors will be protected in accordance with applicable laws.
What advanced notice can I expect before returning to work?	We are working with supervisors and managers to give employees a reasonable amount of time to plan prior to being recalled to the workplace.
Where should face coverings or mask be worn?	DOE is following guidance published by the Centers for Disease Control and Prevention (CDC).
Do I need to wear a mask outside of a building?	DOE is following guidance published by the Centers for Disease Control and Prevention (CDC).

NLP

Normalized Question (no stemming)
collecting personal information upon entry protected
advanced notice expect returning
face covering mask worn
need wear mask outside

Vectorization

BoW model wTFIDF

Sample input question

"Do I need to wear a mask when inside a building?"

NLP

need wear mask inside

*no stemming

Vectorization

BoW model wTFIDF

Similarity Ranking

Q&A Database: Example Ranking (2/2)

- The ranking column specifies how similar is the "input question" to the questions in the database.

Question

"Do I need to wear a mask when inside a building?"

Question	Ranking* (= 1 - Cosine)
Do I need to wear a mask outside of a building?	0.84
Where should face coverings or mask be worn?	0.12
Will DOE be collecting personal information upon building re-entry and, if so, how will it be protected?	0
What advanced notice can I expect before returning to work?	0

*Ranking score of 1.0 would be a perfect match while 0 is no similarity.

- Model accuracy continued to be improved by adding different ways to ask a question to the Q&A database.

Similarity Search: Complex Safety Topics

- DOE Data Analytics and Machine Learning Tools used to analyze ES&H data
  - Search algorithms
  - Data visualization and trending
  - Topic modeling
  - Text clustering

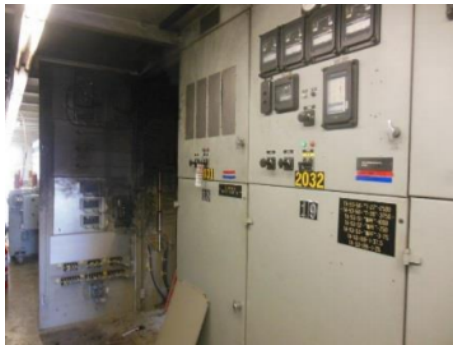
- Leverage the Q&A application to obtain insights in ES&H data and perform more efficient searches

Use Case: Reports related to Oxygen Deficient Atmosphere

- DOE maintains several ES&H databases that are used to:
  - Extract insights from past related events
  - Increase awareness of hazards (e.g., thru safety communications)
- Recent events related to workers accessing oxygen deficient atmosphere (e.g., nitrogen inerted cabinet or room) and passing out or asphyxiating.
- Current tools are limited in how keywords are considered in the searches

Occurrence Reports Search Approaches

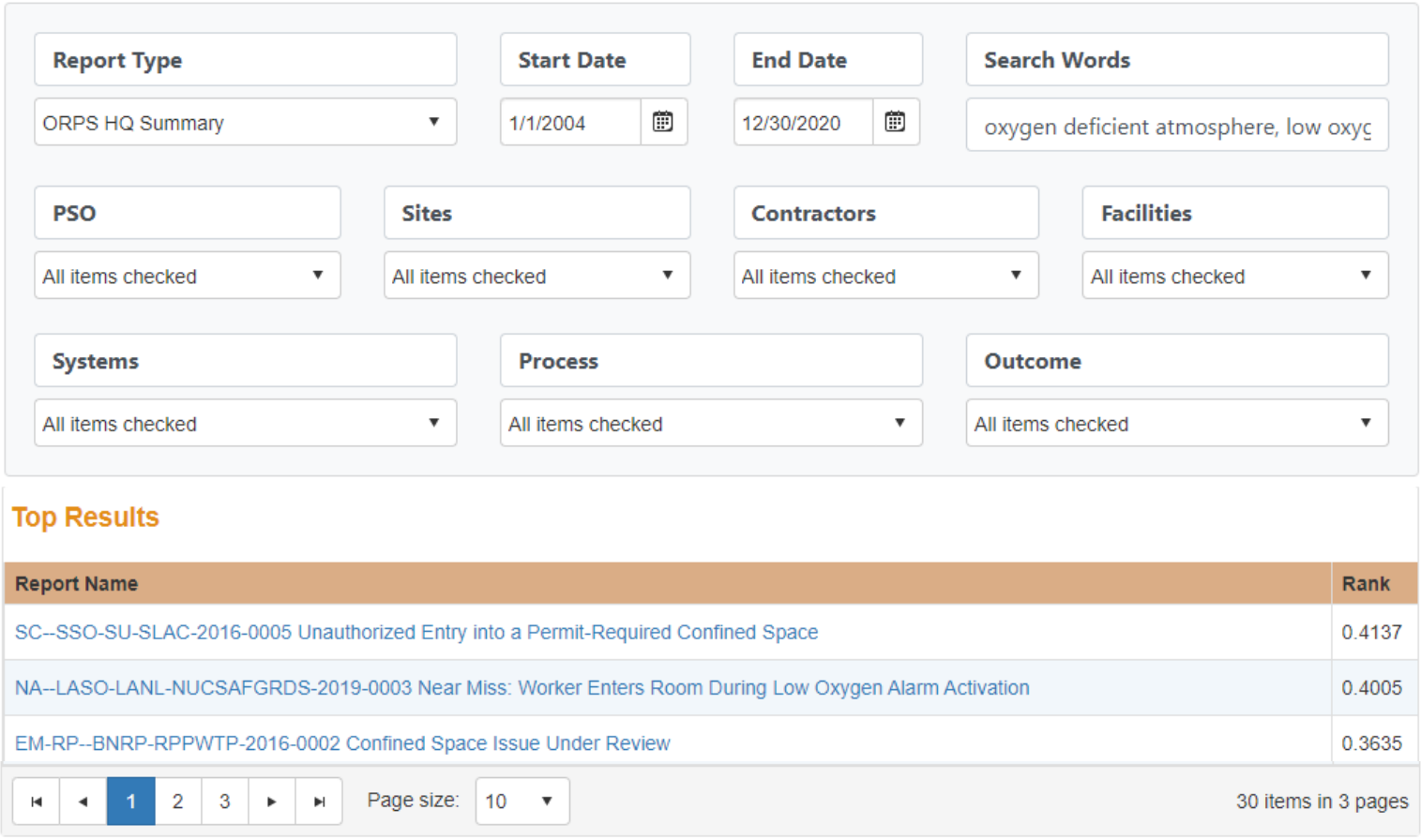
- Topic categorization relies on identified issues of interest (140+ topics are currently tracked)
- Advance information retrieval and search approaches can benefit current systems
- Categorization of occurrence reports help drill down
- Similarity based ranking that relies on the text can be used with multiple keywords or full text of an event description



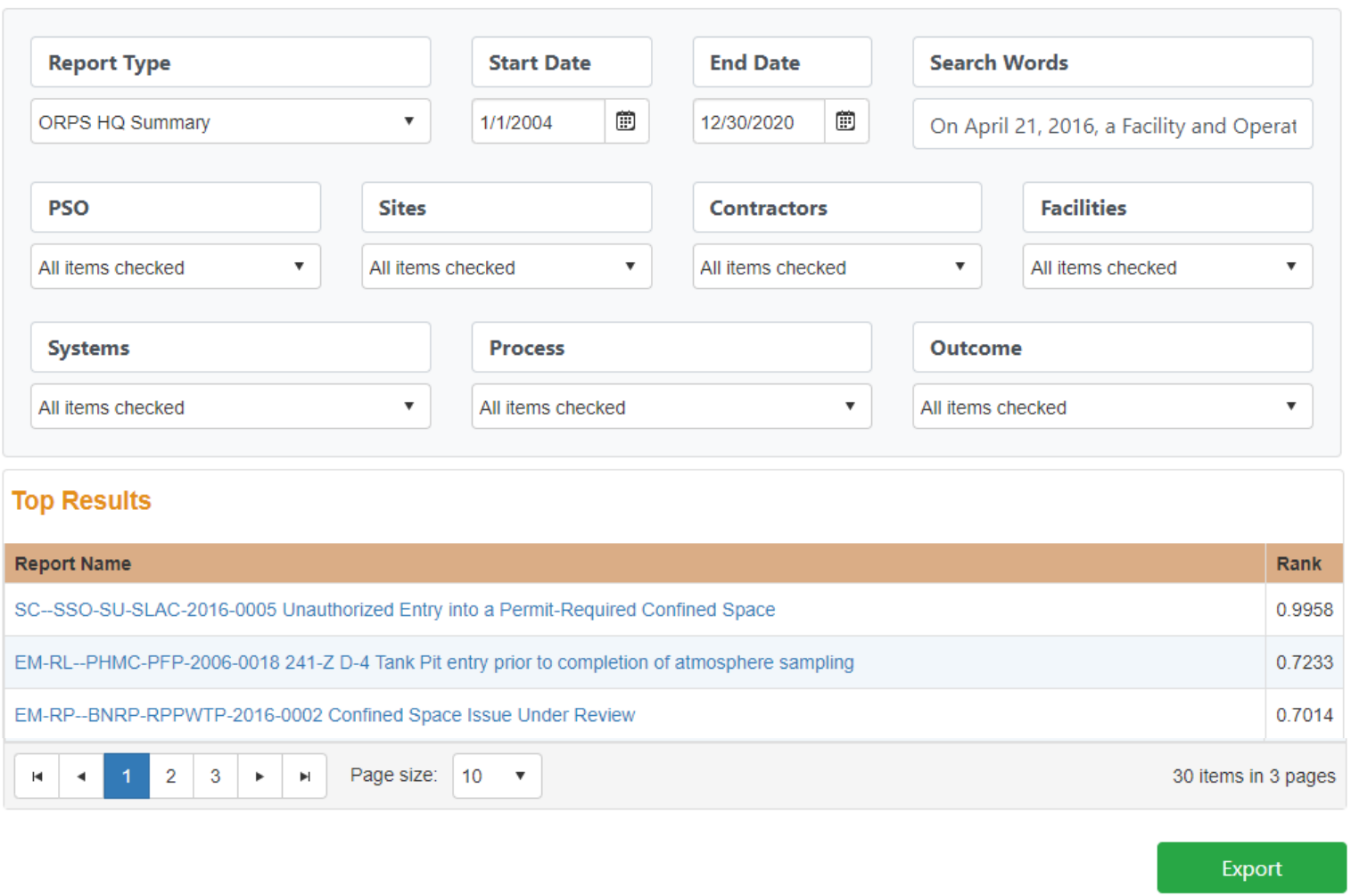
Similarity Search

- Similarity search used to find and rank reports:
  - Using topic keyword search “oxygen deficient atmosphere, low oxygen alarm, nitrogen inert, confined spaces, halon”
  - Using text of a report of interest
- Testing different approaches:
  - Lemmatization
  - Stemming
  - Importance weighting

Similarity Search Dashboard Sample Screen Shot (1/2)



Similarity Search Dashboard
Sample Screen Shot (2/2)



Similarity Search Lessons Learned

- Avoid removing/ignoring words important to the corpus
  - Develop custom stop-words list
  - Do not ignore terms using document frequency parameters
    - max_df =1.0
    - min_df = 0

- Computational costs affected by
  - Size of data
  - Size of BoW model matrix
    - Stop-words
    - N-grams (co-occurring words)
    - Larger values of max_df (up to 1.0)
    - Lower values of min_df

- Stemming is computationally faster than lemmatization and recommended when users don't need to see the normalized text.