

Resource Prediction Using Natural Language Processing

Trey Hathaway
U.S. Nuclear Regulatory Commission
RES/DSA/AAB
August 18, 2021

NRC Data Science and Artificial Intelligence Regulatory Applications Workshops:
Current Topics

Natural Language Processing

- Techniques that allow computers to understand the contents of natural language
 - Allows for the extraction of information and insights from documents
 - Collection of techniques:
 - Rule-based, statistical, or neural

Structured Data

20%

Unstructured Data

PDFS

WORD DOCUMENTS

SPREADSHEETS

PRESENTATIONS

SOCIAL MEDIA POSTS

BOOKS

80%

Use Cases Goals



Apply Natural
Language Processing
techniques to NRC
data and use cases



Demonstrate
Successes

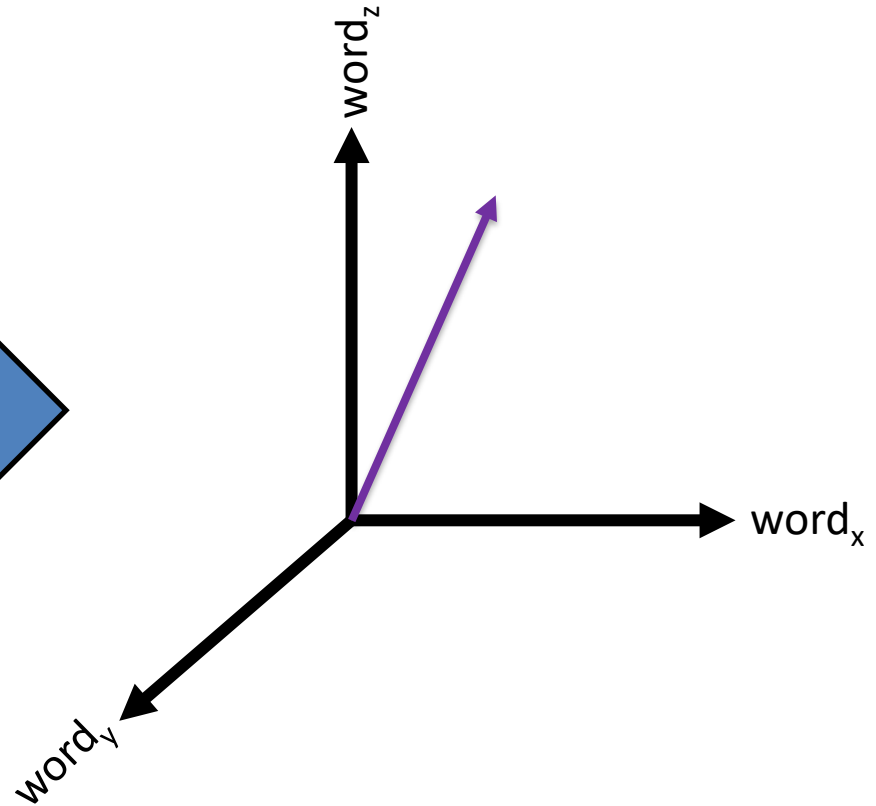
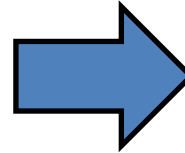
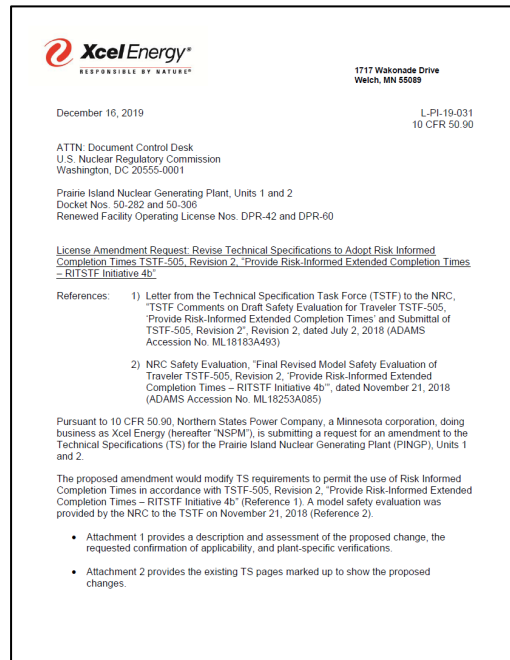
Resource Prediction

- **Challenge:** Deviations between resource estimates to complete a licensing review and the actual hours charged
- **Goal:** Create tool to assist project managers in formulating resource estimates
 - Leverage historical data
 - Find historically similar reviews
- **Method:** Use term frequency-inverse document frequency vectors to represent documents and perform similarity calculations
 - Rank documents based on similarity

Resource Prediction

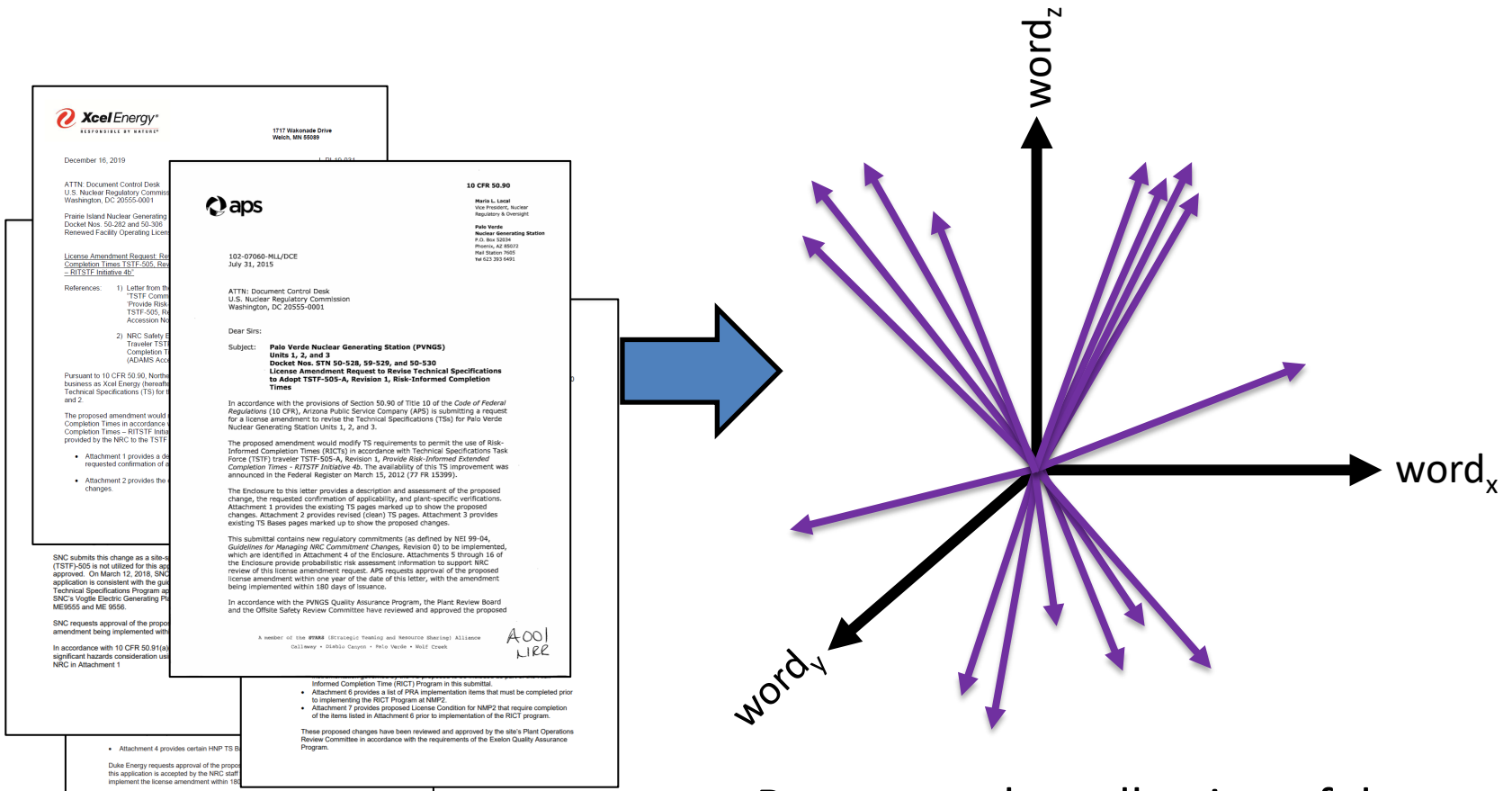
- **Term Frequency-Inverse Document Frequency (tf-idf)**
 - Weighting factor for words
 - Product term frequency and inverse document frequency
- **Term Frequency (tf)**
 - How frequency a word appears in a document
 - Importance of word
- **Inverse document Frequency (idf)**
 - How frequently a word appears in a collection of documents

Term Frequency-Inverse Document Frequency (Vector Representation)



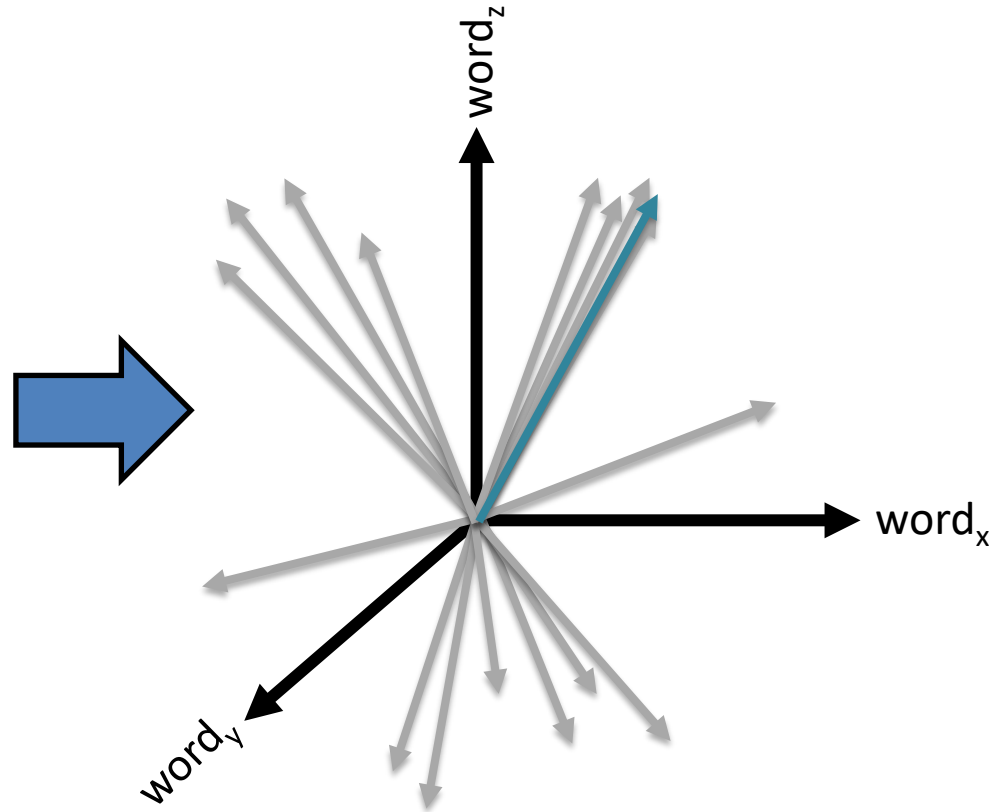
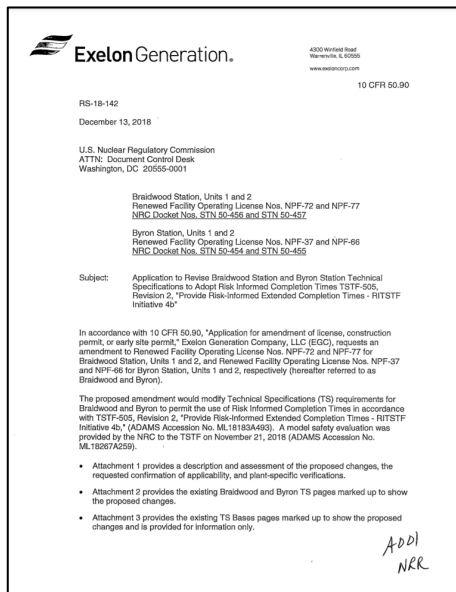
- Represent a document as a vector
 - The vector reflects the word usage in the document
 - The vector will have 1000's of dimensions

Term Frequency-Inverse Document Frequency (Vector Space Corpus)



- Represent the collection of documents as vectors
 - Create a vocabulary of all words used in the collection

Term Frequency-Inverse Document Frequency (Similarity Calculations)



- A new document is converted to a vector based on the vocabulary of the collection of documents
 - The similarity (angle between vectors) is calculated as the dot product between vectors
 - Documents ranked by similarity score

Resource Prediction

Approach

- Acquire historical licensing actions and resource requirements
- Extract text data from pdf files
- Clean data
- Create tf-idf matrix
- Create User Interface
 - Extracts text data
 - Performs similarity calculations

Resource Estimation Tool

U.S. NRC
United States Nuclear Regulatory Commission
Protecting People and the Environment

Resource Estimation Tool

Similarity Screening

Drag and Drop or [Select Files](#)

Similarity: 0.7

Summary

Documents Included: 24
Documents Excluded: 0

Total Hours Charged Summary

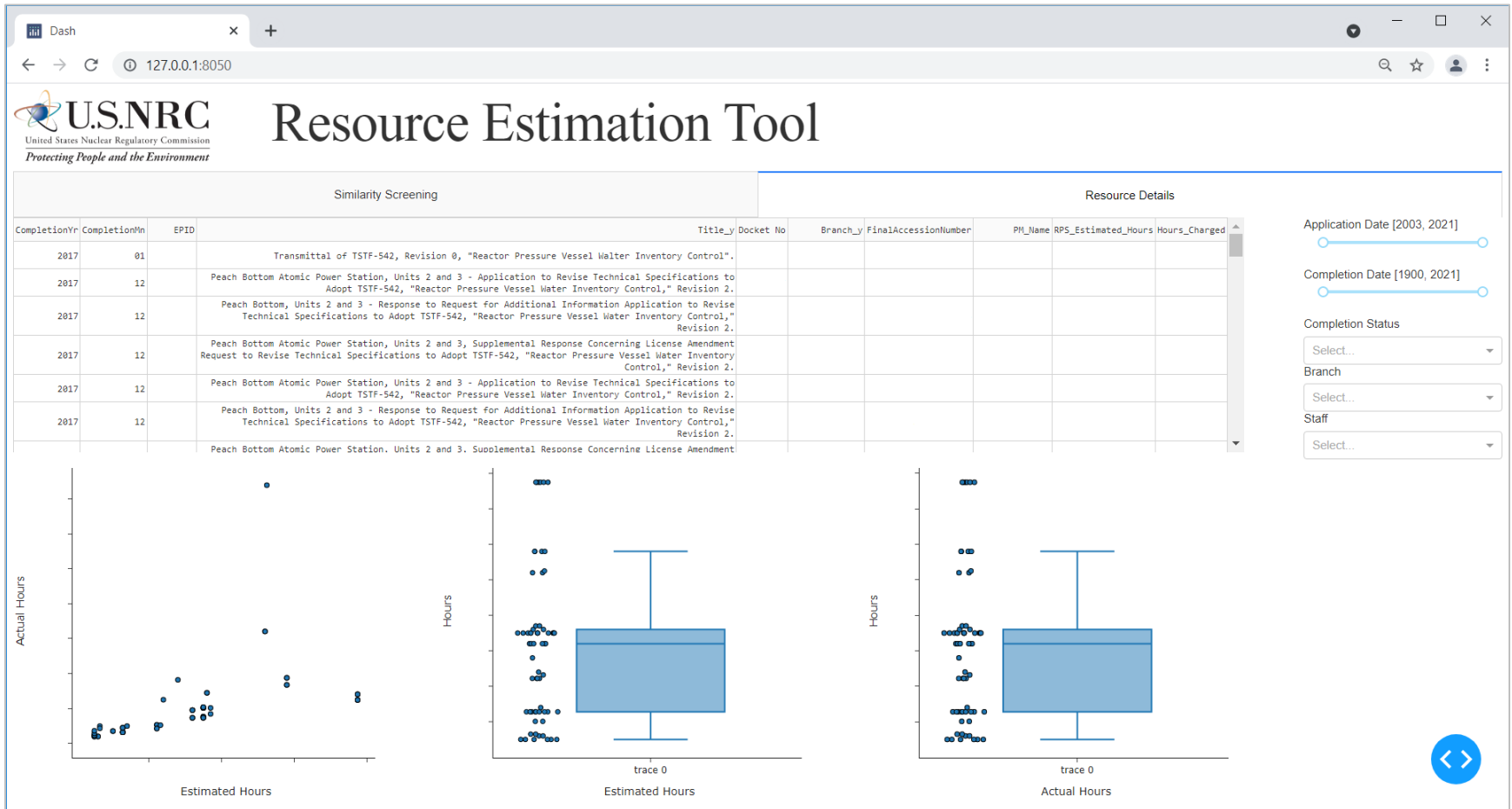
Mean:
Median:
95th Percentile:

Review Duration Summary

Mean:
Median:
95th Percentile:

Accession Number	ProjectId Number	Pages	Author Affiliation	Docket Number	Document Type	Title	Date Received	Total Hours Charged	Review Duration	Branch	Similarity
<input checked="" type="checkbox"/> ML20241A240		109	Southern Nuclear Operating Co, Inc		Letter, License-Application for Facility Operating License (Amend/Renewal) DKT 50	Southern Nuclear Operating Co - Application to Revise Technical Specifications to Adopt TS1F-582, "Reactor Pressure Vessel Water Inventory Control (RPV WIC) Enhancements"				NRR/DORL/LPL21	1
<input checked="" type="checkbox"/> ML20324A090	Package		Exelon Generation Co, LLC		Letter, License-Application for Facility Operating License (Amend/Renewal) DKT 50, Technical Specification, Bases Change, License-Application for Facility Operating License (Amend/Renewal) DKT 50, Technical Specification - Bases Change	Package				NRR/DORL/LPL3	0.88

Resource Estimation Tool



Current Status and Follow-on Work

- Preliminary acceptance testing complete
 - Historical data provides reasonable estimates of required resources and review durations
- NRR/EMBARK and NRR/DORL coordinating to finalize visualizations
- Develop and deploy final User Interface
- Potential Follow-on Work:
 - Search capabilities
 - Predict Branch assignments
 - Predict Standard Review Plan
 - Predict which Regulatory Guide(s) was used for the licensing action

Regulatory Named Entity Recognition

- **Challenge:** Title 10 of the Code of Federal Regulations (CFR), and other regulatory documents, reference sections of 10 CFR
 - Revisions to 10 CFR could impact other sections
- **Goal:** Create a tool to find and extract 10 CFR references from documents
- **Method:** Use Named Entity Recognition (NER) to label text as regulations and extract that text

Named Entity Recognition

(2 CARDINAL) On or before July 26, 1990 DATE , each holder of an operating license for a production or utilization facility in effect on July 27, 1990 DATE , shall submit information in the form of a report as described in 10 CARDINAL CFR 50.75 of this part, indicating how reasonable assurance will be provided that funds will be available to decommission the facility. [21 CARDINAL FR 355, Jan. 19, 1956 DATE , as amended at 35 FR 19660 DATE , Dec. 29, 1970 DATE ; 38 CARDINAL FR 3956, Feb. 9, 1973 DATE ; 45 CARDINAL FR 55408, Aug. 19, 1980 DATE ; 49 CARDINAL FR 35752 WORK_OF_ART , Sept. 12, 1984 DATE ; 53 CARDINAL FR 24049 DATE , June 27, 1988 DATE ; 69 CARDINAL FR 4448 WORK_OF_ART , Jan. 30, 2004 DATE ; 72 CARDINAL FR 49490 CARDINAL , Aug. 28, 2007] DATE 4
Emergency planning zones (EPZs) are discussed in NUREG-0396 DATE , EPA ORG 520/1-78-016, Planning Basis for the Development of State and Local Government Radiological Emergency Response Plans NORP in Support of Light-Water Nuclear Power Plants, December 1978 DATE .

SpaCy Default Entities

(2 CARDINAL) On or before July 26, 1990 DATE , each holder of an operating license for a production or utilization facility in effect on July 27, 1990 DATE , shall submit information in the form of a report as described in 10 CFR 50.75 REG of this part, indicating how reasonable assurance will be provided that funds will be available to decommission the facility. [21 FR 355 FR , Jan. 19, 1956 DATE , as amended at 35 FR 19660 FR , Dec. 29, 1970 DATE ; 38 FR 3956 FR , Feb. 9, 1973 DATE ; 45 FR 55408 FR , Aug. 19, 1980 DATE ; 49 FR 35752 FR , Sept. 12, 1984 DATE ; 53 FR 24049 FR , June 27, 1988 DATE ; 69 FR 4448 FR , Jan. 30, 2004 DATE ; 72 FR 49490 FR , Aug. 28, 2007] DATE 4
Emergency planning zones (EPZs) are discussed in NUREG-0396 NUREG , EPA ORG 520/1-78-016, Planning Basis for the Development of State and Local Government Radiological Emergency Response Plans NORP in Support of Light-Water Nuclear Power Plants, December 1978 DATE .

Addition of NRC Specific Language Patterns

- **Used Python package Spacy**

10 CFR Reference Identification Tool

Dash 127.0.0.1:8050 Guest (3) Update

Natural Language Processing Tool

Instructions for Sunburst Chart:

Please click on a slice in the outer ring to choose a 10 CFR part. Then click a slice in the updated outer ring to choose a section within that part. The part/section will be in the middle with an outer ring for the locations that references the middle. Click on one of those slices and scroll down to see the output that contains a hyperlink to the full documentation as well as sentences/paragraphs of interest.

Instructions for Input Box:

Please type in a part and section to the input box in the format: XX.XX()(), with the parentheses being optional. If interested in a part with no specific section in mind, please type in the part number in the format: X.0. Once you click enter or click outside of the input box, scroll down to see the output. It will include the locations the part/section is referenced within 10 CFR, along with a link to the full documentation, and sentences/paragraphs of interest.

10 CFR

10 CFR Reference Identification Tool

Dash 127.0.0.1:8050 Guest (3) Update

Please type in a part and section to the input box in the format: XX.XX()(), with the parentheses being optional. If interested in a part with no specific section in mind, please type in the part number in the format: X.0. Once you click enter or click outside of the input box, scroll down to see the output. It will include the locations the part/section is referenced within 10 CFR, along with a link to the full documentation, and sentences/paragraphs of interest.

XX.XX() ()

Output from the Sunburst Chart:

Click the link below to see the full documentation:

[10 CFR 50.58](#)

Below is a snippet with the reference of interest (each paragraph is separate):

(a) Each application for a construction permit or an operating license for a facility which is of a type described in § 50.21(b) or § 50.22, or for a testing facility, shall be referred to the Advisory Committee on Reactor Safeguards for a review and report. An application for an amendment to such a construction permit or operating license may be referred to the Advisory Committee on Reactor Safeguards for review and report. Any report shall be made part of the record of the application and available to the public, except to the extent that security classification prevents disclosure.

(b)(1) The Commission will hold a hearing after at least 30-days' notice and publication once in the FEDERAL REGISTER on each application for a construction permit for a production or utilization facility which is of a type described in § 50.21(b) or § 50.22, or for a testing facility.

(5) The Commission will use the standards in § 50.92 to determine whether a significant hazards consideration is presented by an amendment to an operating license for a facility of the type described in § 50.21(b) or § 50.22, or which is a testing facility, and may make the amendment immediately effective, notwithstanding the pendency before it of a request for a hearing from any person, in advance of the holding and completion of any required hearing, where it has determined that no significant hazards consideration is involved.

Navigation arrows: <>

10 CFR Reference Identification Tool

50.82(a)(1)

10 CFR

Output from the input box:

50.4: <https://www.nrc.gov/reading-rm/doc-collections/cfr/part050/part050-0004.html>

(8) Certification of permanent cessation of operations. The licensee's certification of permanent cessation of operations, under § 50.82(a)(1), must state the date on which operations have ceased or will cease, and must be submitted to the NRC's Document Control Desk. This submission must be under oath or affirmation.

(9) Certification of permanent fuel removal. The licensee's certification of permanent fuel removal, under § 50.82(a)(1), must state the date on which the fuel was removed from the reactor vessel and the disposition of the fuel, and must be submitted to the NRC's Document Control Desk. This submission must be under oath or affirmation.

50.36: <https://www.nrc.gov/reading-rm/doc-collections/cfr/part050/part050-0036.html>

(6) Decommissioning. This paragraph applies only to nuclear power reactor facilities that have submitted the certifications required by § 50.82(a)(1) and to non-power reactor facilities which are not authorized to operate. Technical specifications involving safety limits, limiting safety system settings, and limiting control system settings; limiting conditions for operation; surveillance requirements; design features; and administrative controls will be developed on a case-by-case basis.

50.36b: <https://www.nrc.gov/reading-rm/doc-collections/cfr/part050/part050-0036b.html>

(b) Each license authorizing operation of a production or utilization facility, including a combined license under part 52 of this chapter, and each license for a nuclear power reactor facility for which the certification of permanent cessation of operations required under § 50.82(a)(1) or § 52.110(a) of this chapter has been submitted, which is of a type described in § 50.21(b)(2) or (3) or § 50.22 or is a testing facility, may include conditions to protect the environment during operation and decommissioning. These conditions are to be set out in an attachment to the license which is incorporated in and made a part of the license. These conditions will be derived from information contained in the environmental report or the supplement to the environmental report submitted pursuant to §§ 51.50 and 51.53 of this chapter as analyzed and evaluated in the NRC record of decision, and will identify the obligations of the licensee in the environmental area, including, as appropriate, requirements for reporting and keeping records of environmental data, and any conditions and monitoring requirement for the protection of the nonaquatic environment.

50.44: <https://www.nrc.gov/reading-rm/doc-collections/cfr/part050/part050-0044.html>

(b) Requirements for currently-licensed reactors. Each boiling or pressurized water nuclear power reactor with an operating license on October 16, 2003, except for those facilities for which the certifications required under § 50.82(a)(1) have been submitted, must comply with the following requirements, as applicable:

Conclusions

- Natural Language Processing is a powerful tool to leverage unstructured data in historical documents
- Deploying these tools would increase efficiency of staff by reducing time required for manual searches
 - Staff can leverage historical data in informing decisions