



Webinar

RangeDocs Searchable Science for Rangeland Management

August 21, 2024



Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.
This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

P



Jason Karl, University of Idaho

Overview of RangeDocs



Matt King, University of Arizona

RangeDocs WebApp



Kristina Riemer & J.D. Gibbs, University of Arizona

Machine Learning & AI



Question & Answer Session

òŒEd'ŽÒ?ĒΦúĒpÒdĒÒpΦdÍτΦÒŒZĒYŒĒĀŒÒ

THE CHALLENGE

Technical references and scientific papers are often long and dense. With limited time and energy to read literature, finding specific information can feel impossible.

THE SOLUTION

Range Docs is an online tool designed to search technical references and other literature at the page level. Rangeland experts have read the documents and labelled common rangeland terms where they appear in the documents. Science is now searchable.



YCHONOCIZTODUZO

Unique attributes of RangeDocs:

1. Leverages a rangeland-specific thesaurus of terms harmonized from multiple sources
2. Contains the most relevant and useful resources as identified by agency staff and Extension specialists
3. Uses paragraph-level annotation of concepts to get users quickly to the most relevant information

Core Project Components

Grazingland Thesaurus

- Coordinate/update existing glossaries as a basis for tagging and finding grazingland reference materials

Content Selection

- Work with the Rangelands Partnership, NRCS and others to identify content

Content Indexing

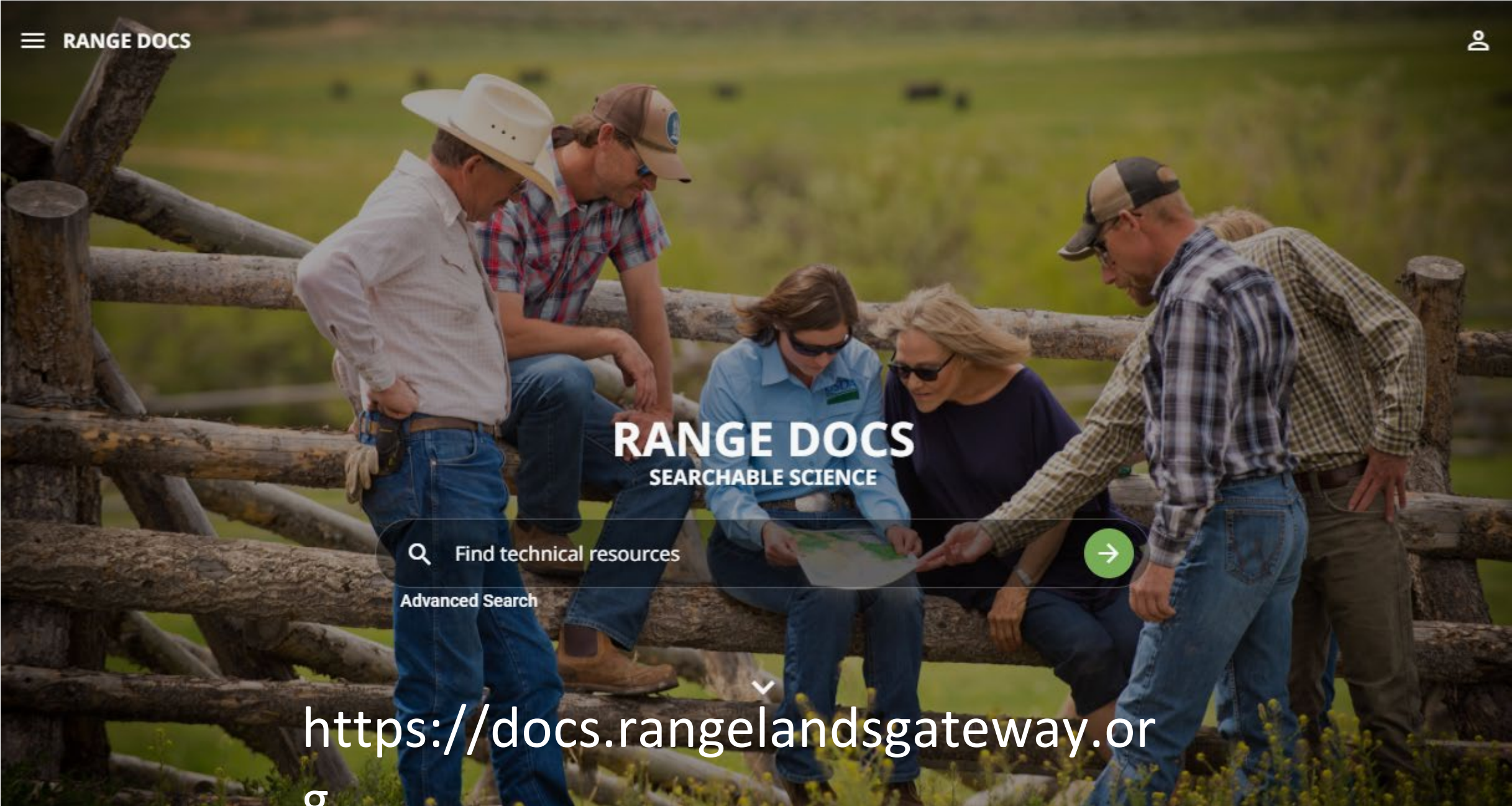
- Tag documents and document sections with terms from the thesaurus for improved search results

Application Development

- Develop web and mobile interfaces for the system within the Global Rangelands platform

Outreach and Training

- Work with the Rangelands Partnership and other partners to engage users in helping develop and implement the system



RANGE DOCS

SEARCHABLE SCIENCE

🔍 Find technical resources

Advanced Search



https://docs.rangelandsgateway.org



òŒd'žò?îê?îê→đòd!!
! d'd'í !!Œ!p d'

☰ RANGE DOCS 👤

MEASURING AND MONITORING PLANT POPULATIONS

This technical reference applies to monitoring situations involving a single plant species, such as an indicator species, key species, or weed....

🚨 43

MONITORING MANUAL FOR GRASSLAND, SHRUBLAND AND SAVANNA ECOSYSTEMS: VOLUME 2 - DESIGN, SUPPLEMENTARY METHODS AND INTERPRETATION

The Monitoring Manual for Grassland, Shrubland and Savanna Ecosystems is divided into two volumes: This two-volume document is intended to assist...

🚨 3

MONITORING MANUAL FOR GRASSLAND, SHRUBLAND AND SAVANNA ECOSYSTEMS:

ANNOTATING

Page Access

START ANNOTATING

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64
65	66	67	68	69	70	71	72
73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88

ISSUES

17 "Adaptive Management" is not currently a valid tag.

18 "Resource Monitoring" is not currently a valid tag.
"Adaptive Management" is not currently a valid tag.

22 "Baseline Monitoring" is not currently a valid tag.
"Adaptive Management" is not currently a valid tag.

23 "Long-term Ecological Monitoring" is not currently a valid tag.

"Monitoring Plan" is not

- Documents are annotated by rangeland professionals using concepts from the thesaurus

ðŒd'ZÒ?ËŒ?Ë → ðŒd!!
! d'd' !Œ!p d'

- Concepts are annotated at the page and paragraph level
- Allows search to direct user to specific parts of a document

The screenshot shows a document viewer interface. The main document area has several paragraphs of text with yellow highlights. The sidebar on the right shows a list of search results for the term 'Range Docs', with each result corresponding to a highlighted section in the document. The interface includes a search bar at the top, a document title 'RANGE DOCS', and a sidebar with a search icon and a list of results.

(b) Establishing management objectives

Management objectives are developed and determined with the landowner during the planning process. All inventory and other necessary information for the development of objectives and the application of the grazing management are gathered during the planning process. The objectives of the landowner and those of the NRCS do not need to be the same, but they must be compatible. The management objective must meet the needs of the landowner, the resources, and the grazing animals.

(c) Determining treatment alternatives

The NRCS conservationist will use information from the ecological site description, trend determinations, similar...

This stage of the conservation involves the following steps:

- Inventory the present plant determine annual produc
- Identify from the ecologi desired plant community manager's goals and the r
- Determine what changes (determine trend).
- Compute similarity index to the desired plant comm
- Determine how the ecolc site are functioning (rang nations).
- Determine what conserva tives and resulting resour will achieve or maintain

òœEd'ŽÒ?ĚΦúÒœEŽÊYđÒΦ→!Φ

Search results direct users to the page and paragraph level

Annotated documents are prioritized (show up first)

Accessed documents are cached for offline access

The screenshot shows a search interface with a search bar containing the text "stocking rate". Below the search bar, it indicates "RESULTS 1 - 16 of 1182". The first result is titled "Estimating Initial Stocking Rates" by Dan Ogle and Brendan Brazee, published in 2009 and consisting of 39 pages. The abstract of the document is visible, discussing the importance of stocking rates for animal performance and land health. A preview of the document content is shown at the bottom, with the search term "stocking rate" highlighted in green. The preview includes the text: "ESTIMATING INITIAL STOCKING RATES Dan Ogle, Plant Materials Specialist, NRCS, Boise, IDBrendan Brazee ... Stocking rate has the largest impact on animal performance and the health of the forage resource of all ... composition over the long term xPlant physiology xProfitability of the operation Establishing a proper stocking ... Factors that affect stocking rate include the animal species, class of livestock (dry cow, lactating ... With this in mind, setting the appropriate initial stocking rate consists of determining (1) how much".

Βύττε!!òφ!!
ò ρα'žò? í êφ

Local Extension documents

Conservation Practice Standards &
2022 NRPH

User Guide

Support Videos



RangeDocs

Searchable Science for Rangeland Management

Project Team

Jason Karl
Amber Dalke
Sean DiStefano
Barb Hutchinson
Jeremy Kenyon
Matt King
Karen Launchbaugh
Sheila Merrigan
Jeanne Pfander
Matt Rahr
George Ruyle
Eric Winford
Retta Bruegger
Mark Thorne
J.D. Gibbs
Kristina Riemer



Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.

This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

Additional assistance provided by members of the Rangelands Partnership and the Altar Valley Conservation Alliance.

P



Jason Karl, University of Idaho

Overview of RangeDocs



Matt King, University of Arizona

RangeDocs WebApp



Kristina Riemer & J.D. Gibbs, University of Arizona

Machine Learning & AI



Question & Answer Session



òĈEd'ŽÒ?ĚΦÒÂ!üü
! ž—ĤĚČ ů òĈĈ—đĩ òž!!Ýò Ýí í ĩ

What problems were we trying to solve?

- Search against a repository of PDFs
- Provide relevant search results
 - Not only at the file level, but at the page level
- Annotate PDF text selections with new rangelands common terms
- Use the annotations as a mechanism to boost relevancy of pages



òĈEd'ŽÒ?ĚΦÒÂ!üü
! ž—ĤĚč ůòĈĈ—đĩ òž!!Ýò ýí í ĩ

Off the shelf products and search engines only get us so far

We didn't want to reinvent the wheel...

... but rather make a new wheel out of parts of other wheels.



RangeDocs WebApp

A quick peek under the hood

1. Page level queries

Give internal pages of large PDFs their own relevancy

- Provide users direct access to internal pages of PDFs without losing context

2. Annotations

Tagging areas of text with common Rangeland terms

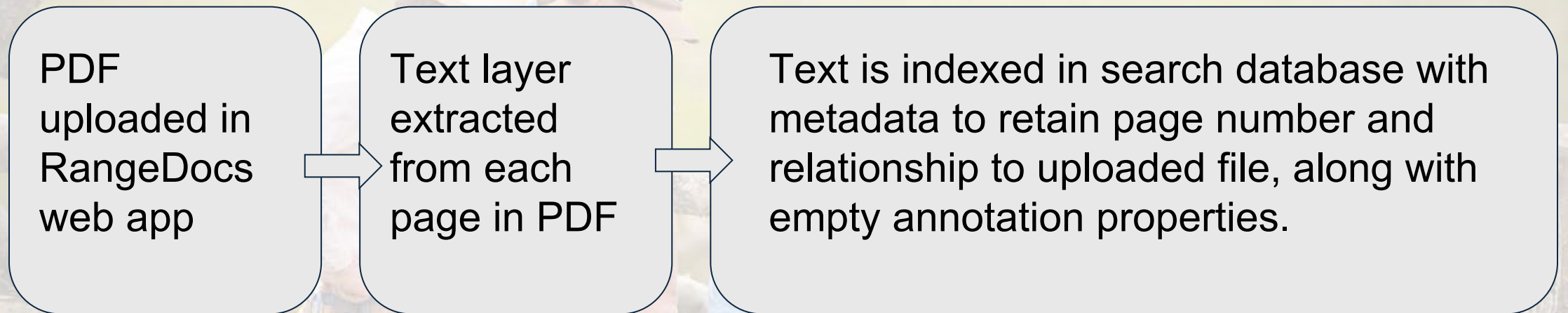
- Term not always explicitly stated in the page or paragraph
- Reinforces meaning of page/text by association of terms



RangeDocs WebApp

A quick peek under the hood

Page level queries



- This gives each page its own relevancy score when being searched
- The score is boosted by overlapping relevancy with the original file's metadata (e.g. title, authors, abstract, tags, etc)



RangeDocs WebApp

A quick peek under the hood

Annotations: Powered by [Hypothes.is](https://hypothes.is)

Online open source web annotation tool

- Third party code embedded in RangeDocs to annotate any text selection
 - Text selections can be uniquely identified by their placement on a webpage, PDF fingerprints, and web page URL
- Hypothes.is provides private groups that allow only elevated users in RangeDocs to submit valid annotations



RangeDocs WebApp

A quick peek under the hood

Annotations: Hypothes.is (cont'd)

Annotator selects paragraph in PDF in RangeDocs

Text selection tagged with term from glossary

Annotations are saved in Hypothesi.is' system. RangeDocs uses Hypothes.is API to harvest / synch annotations to the page annotation properties.

- Additional boost algorithms are given to pages where search has overlapping relevancy with page annotations



RangeDocs WebApp

A quick peek under the hood

Collections: new virtual documents

- Leveraging the output of Page Level Queries, PDFs uploaded into RangeDocs are stored as individual documents
- Collections allow any user to stitch together different pages from across all resources in RangeDocs
 - Colate like topics into one place
- Private vs Curated
 - View virtually in the browser or download as a new PDFs



RangeDocs

Searchable Science for Rangeland Management

Project Team

Jason Karl
Amber Dalke
Sean DiStefano
Barb Hutchinson
Jeremy Kenyon
Matt King
Karen Launchbaugh
Sheila Merrigan
Jeanne Pfander
Matt Rahr
George Ruyle
Eric Winford
Retta Bruegger
Mark Thorne
J.D. Gibbs
Kristina Riemer



Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.

This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

Additional assistance provided by members of the Rangelands Partnership and the Altar Valley Conservation Alliance.

P



Jason Karl, University of Idaho

Overview of RangeDocs



Matt King, University of Arizona

RangeDocs WebApp



Kristina Riemer & J.D. Gibbs, University of Arizona

Machine Learning & AI



Question & Answer Session



RESEARCH
u OCEYpd'O o OCEZpd'Z

OBJECTIVE

Improve power and extensibility of Rangedocs system with modern modeling tools.

1. Train language-based models on set of annotated documents
2. Incorporate into Rangedocs Gateway search
3. Automate the annotations of new PDFs to need little or no human assistance



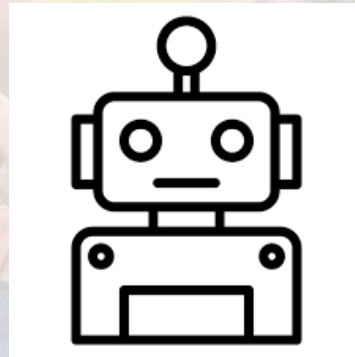
የግንባታ ስራ ለማረጋገጥ
a YÒZÒ-Ò-Ò-ÀÒÒd'

Created *machine learning models* to classify new documents using annotated documents for training



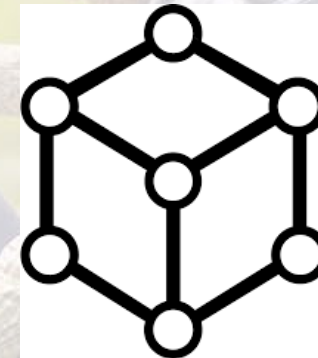
Text & tags

+



Algorithm

=



Model for each tag



Caltech
a YÒZÒ-Ô-Ò-ÀÒÒd'

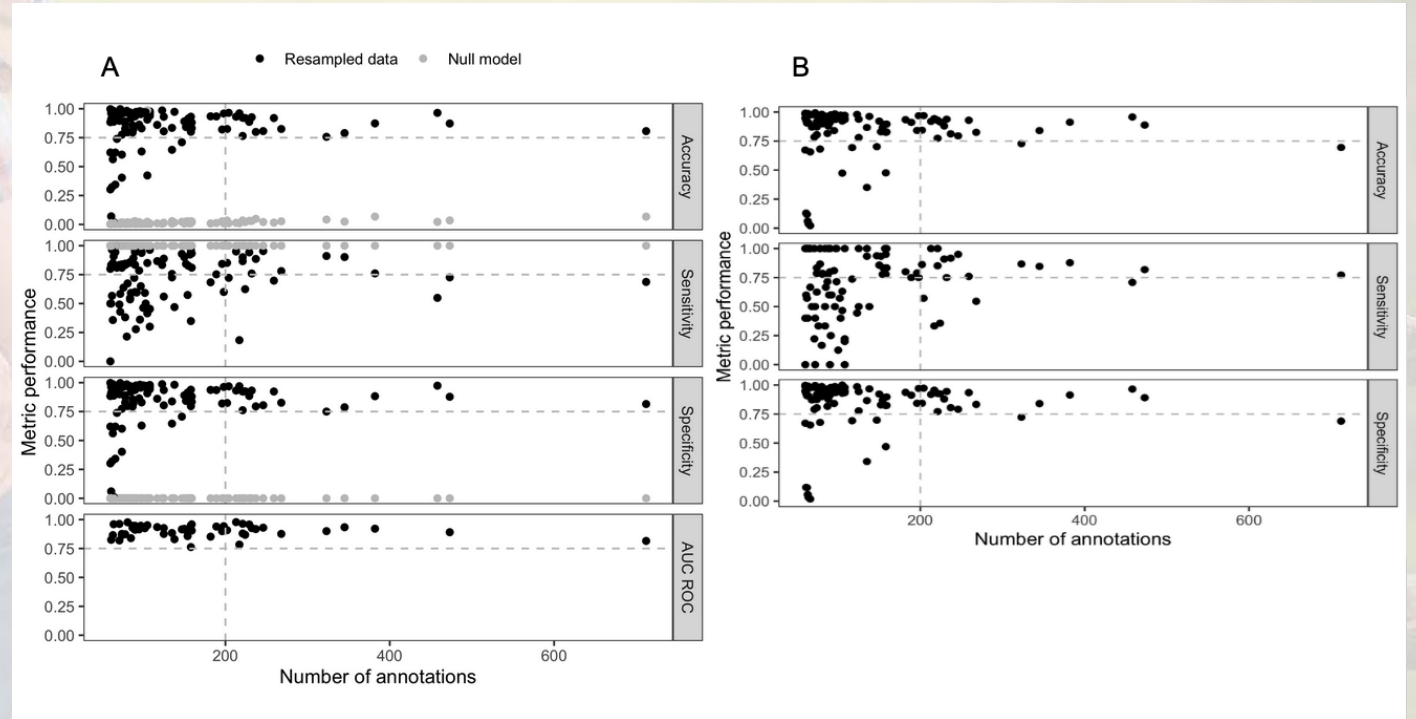
Summary

- Chose classification model
 - Support vector machine w/ polynomial kernel
- Tokenized documents
- Fit model to training data
- Evaluated performance w/ standard metrics
- Documented in code repository
 - <https://bitbucket.org/cals-cct/docs-classification-model/src/main/>



የግንባታ ስራ ለማረጋገጥ
a ሃዕዓዕ ጥዕዕ ለሰዕዕ

- Results
 - Can only reliably predict correct thesaurus terms if tagged at least 200 times in text
- Most terms were annotated much fewer times, so ML models weren't as useful
- Then LLMs came on the scene!

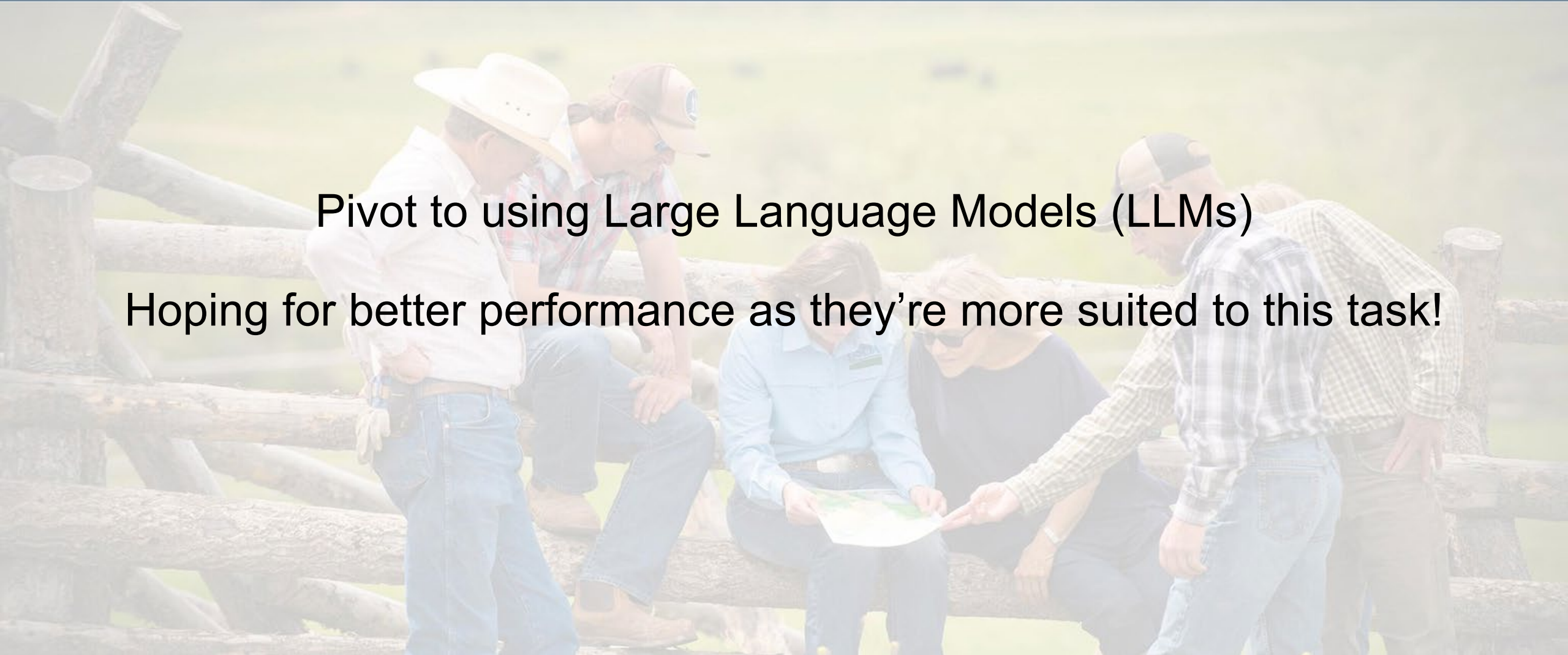




ᐃᑦᑲᑦᑲᑦ
a ᐃᑦᑲᑦᑲᑦ ᐃᑦᑲᑦᑲᑦ

Pivot to using Large Language Models (LLMs)

Hoping for better performance as they're more suited to this task!





RangeDocsML

What are LLMs?

- Large Language Models (LLMs) are advanced AI models designed to understand, generate, and interact with human language in a sophisticated manner
 - Trained on vast amounts of text data from diverse sources to develop a nuanced understanding of language
 - Can perform a wide range of language-related tasks, including text generation, translation, summarization, and question-answering

- Examples:

- ChatGPT (OpenAI)
- Gemini (Google)
- Claude (Anthropic)
- Llama (Meta)



OpenAI

Gemini



Claude

BY ANTHROPIC

∞ Meta



ŒŒŒŒ

ooü

| -òž|pòT

Tokenize thesaurus terms

Store and retrieve data as vectors in a vector database

- Provides efficient search capabilities for fast retrieval of relevant documents

œŕçŕ

ooü

| -òz|pòt

Use pretrained models and fit to our use case

Chose dense passage retrieval models

- Enables asking a question using terms of interest and getting relevant document text back
- Compare multiple models
- Tune hyperparameters

RangeDocs LLM Overview

Evaluate model performance

Determine how well model does using standard metrics

Čp
ů ž Žž

Tokenize thesaurus terms



Tokenize range documents



Use pretrained models and fit to our use case

Evaluate model performance

òŒEd'ŽÒ?ÍÊΦuoιϑ ϕŽŒEpd'pd'Ž

OpenSource Models (HuggingFace)

- Testing models fine-tuned for retrieval to see which best fits our use case
- Evaluate model performance with metrics and adjust with Hyperparameter Tuning for optimal results

Vector Database

- Search Optimization: Provides efficient search capabilities for fast retrieval of relevant documents
- chromadb or pinecone.io

Fine Tuning

- Training with Jetstream2 GPU Allocation
- Flexible codebase for multiple collaborators, shared learning and quick coordinations
 - Using Tools such as: Python, Anaconda, JupyterLabs/Notebooks, Bitbucket/GitHub

User Testing

- A/B Testing within WebApp

òŒEd'ŽÒ?ĖΦuo?U→Z!YÒZòÒŒEjpd'Ž

- **Infrastructure**

- Portable learning environments for hands-on computational instruction: Using container- and cloud-based technology to teach data science – [ResearchGate.net](https://www.researchgate.net)

- **Task-Specific Models**

- [HuggingFace.co](https://huggingface.co) provides Open-Source ML infrastructure, models pre-trained on tasks, fine-tuning pipelines for transfer learning.
- Models like `msmarco` ([Sbert docs](https://www.sbert.net/docs)) are fine-tuned specifically for retrieval tasks, improving accuracy.

- **Chunking Strategies**

- Breaking down large pieces of text into smaller segments, optimizing the relevance results from a vector database – [Pinecone.io](https://pinecone.io)

- **Dense Passage Retrieval (DPR)**

- High Accuracy model excelling at retrieving passages by creating dense representations of queries and passages.
- Dual Encoder Architecture: Uses separate encoders for queries and passages, allowing efficient indexing and retrieval, re-ranking and fine-tuning to boost retrieval performance.
- DPR can utilize models that that make it easier to retrieve relevant documents based on semantic similarity, such as those from **Sentence Transformers**, but it employs its own architecture for retrieval tasks

- **Retrieval-Augmented Generation (“RAG”)**

- Combine the powers of pretrained DPR and sequence-to-sequence models
- Retrieves documents, passes to a fine-tuned seq2seq model, then generates outputs, allowing both retrieval and generation to adapt to downstream tasks.



RangeDocs

Searchable Science for Rangeland Management

Project Team

Jason Karl
Amber Dalke
Sean DiStefano
Barb Hutchinson
Jeremy Kenyon
Matt King
Karen Launchbaugh
Sheila Merrigan
Jeanne Pfander
Matt Rahr
George Ruyle
Eric Winford
Retta Bruegger
Mark Thorne
J.D. Gibbs
Kristina Riemer



Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005.

This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.

Additional assistance provided by members of the Rangelands Partnership and the Altar Valley Conservation Alliance.



RangeDocs

Searchable Science for Rangeland Management

docs.rangelandsgateway.org



Funding provided by a USDA Natural Resources Conservation Service Conservation Innovation Grant NR193A750008G003 & NR233A750008C005. This work is supported by the Renewable Resources Extension Act Program [award no. 2021-46401-34740] from the USDA National Institute of Food and Agriculture.