Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

Improving Electronic Logbook Searches Using Natural Language Processing

Jennefer Maldonado

Collider Accelerator Department, Brookhaven National Laboratory

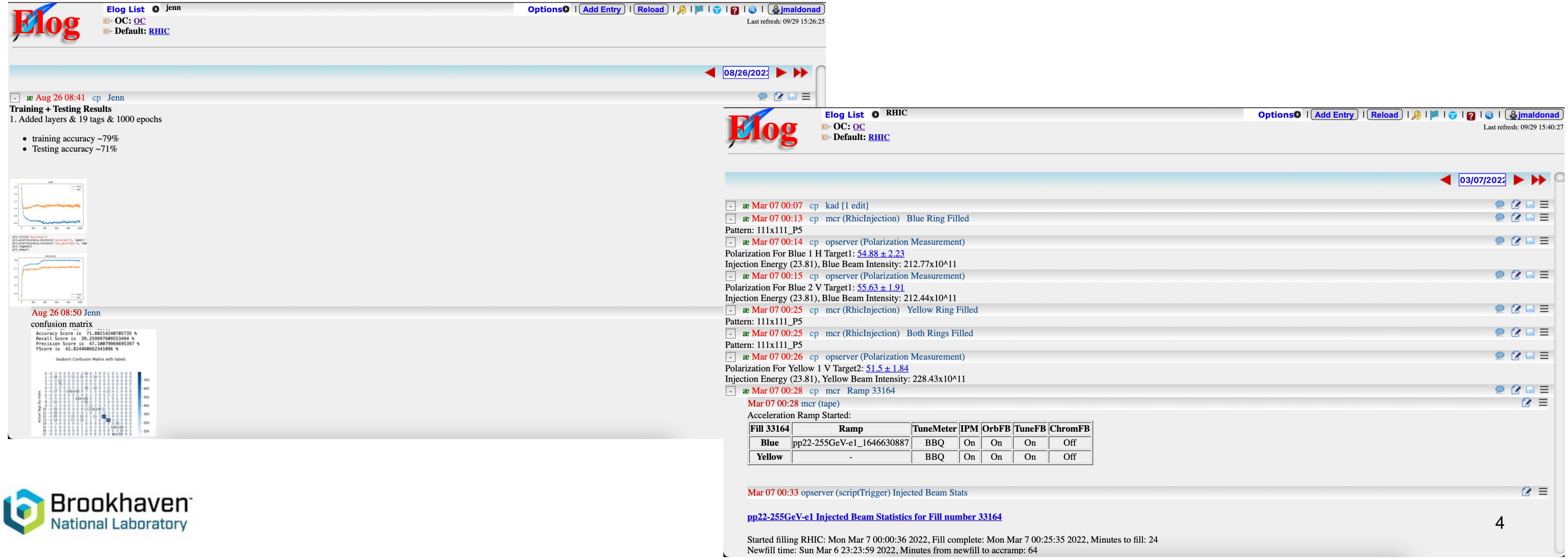
jmaldonad@bnl.gov

@BrookhavenLab

Introduction

What is the elog system?

- The electronic logbook (elog) system is used to record information ranging from meeting notes, to do lists, and critical operations



jenn

OC: OC

Default: RHIC

Last refresh: 09/29 15:26:25

æ Aug 26 08:41 cp Jenn

Training + Testing Results

1. Added layers & 19 tags & 1000 epochs

- training accuracy ~79%
- Testing accuracy ~71%

```
plt.title('Accuracy')
plt.plot(history.history['accuracy'], label='
plt.plot(history.history['val_accuracy'], lab
plt.legend()
plt.show()
```

Aug 26 08:50 Jenn

confusion matrix

```
Accuracy Score is  71.80214240705735 %
Recall Score is  39.259997609553494 %
Precision Score is  47.10079969695397 %
FScore is  42.824460662341096 %
```

RHIC

OC: OC

Default: RHIC

Last refresh: 09/29 15:40:27

æ Mar 07 00:07 cp kad [1 edit]

æ Mar 07 00:13 cp mcr (RhicInjection) Blue Ring Filled

Pattern: 111x111_P5

æ Mar 07 00:14 cp opserver (Polarization Measurement)

Polarization For Blue 1 H Target1: 54.88 ± 2.23

Injection Energy (23.81), Blue Beam Intensity: 212.77x10^11

æ Mar 07 00:15 cp opserver (Polarization Measurement)

Polarization For Blue 2 V Target1: 55.63 ± 1.91

Injection Energy (23.81), Blue Beam Intensity: 212.44x10^11

æ Mar 07 00:25 cp mcr (RhicInjection) Yellow Ring Filled

Pattern: 111x111_P5

æ Mar 07 00:25 cp mcr (RhicInjection) Both Rings Filled

Pattern: 111x111_P5

æ Mar 07 00:26 cp opserver (Polarization Measurement)

Polarization For Yellow 1 V Target2: 51.5 ± 1.84

Injection Energy (23.81), Yellow Beam Intensity: 228.43x10^11

æ Mar 07 00:28 cp mcr Ramp 33164

Mar 07 00:28 mcr (tape)

Acceleration Ramp Started:

Fill 33164	Ramp	TuneMeter	IPM	OrbFB	TuneFB	ChromFB
Blue	pp22-255GeV-e1_1646630887	BBQ	On	On	On	Off
Yellow	-	BBQ	On	On	On	Off

Mar 07 00:33 opserver (scriptTrigger) Injected Beam Stats

pp22-255GeV-e1 Injected Beam Statistics for Fill number 33164

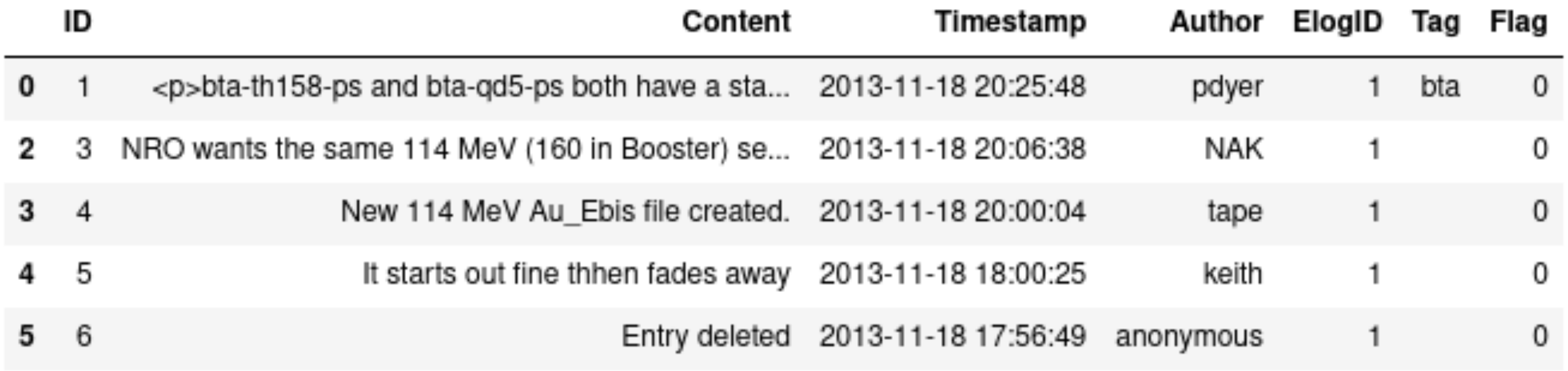
Started filling RHIC: Mon Mar 7 00:00:36 2022, Fill complete: Mon Mar 7 00:25:35 2022, Minutes to fill: 24

Newfill time: Sun Mar 6 23:23:59 2022, Minutes from newfill to accramp: 64

Motivation

- The search feature only provides exactly what a user enters, what if there are other entries that do not include those exact words BUT also are related to these things
- Allow for more catering to search filters like date, logbook, etc.
- Eventually provide custom sets of entries based on users' interactions with the system
- Possibly eliminate the need for manual searching

Data

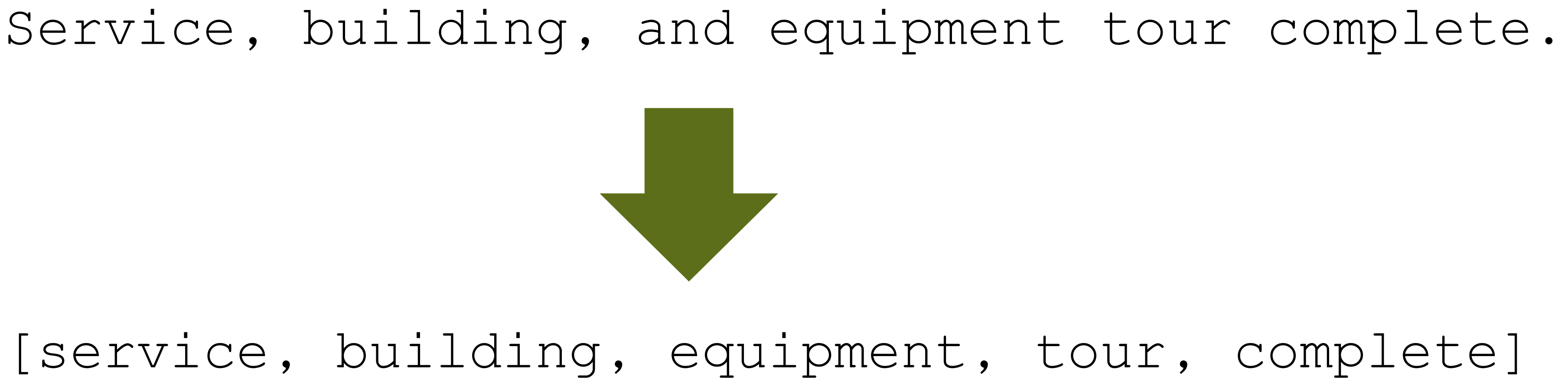


	ID	Content	Timestamp	Author	ElogID	Tag	Flag
0	1	<p>bta-th158-ps and bta-qd5-ps both have a sta...	2013-11-18 20:25:48	pdyer	1	bta	0
2	3	NRO wants the same 114 MeV (160 in Booster) se...	2013-11-18 20:06:38	NAK	1		0
3	4	New 114 MeV Au_Ebis file created.	2013-11-18 20:00:04	tape	1		0
4	5	It starts out fine thhen fades away	2013-11-18 18:00:25	keith	1		0
5	6	Entry deleted	2013-11-18 17:56:49	anonymous	1		0

- The database includes whether entries are a comment, what book they are in, and time entered
- All elog data is stored in a MySQL database

Data Processing

- Remove links, numbers, tokenize, lemmatize, lowercase, remove punctuation, also remove any entries with no content in them



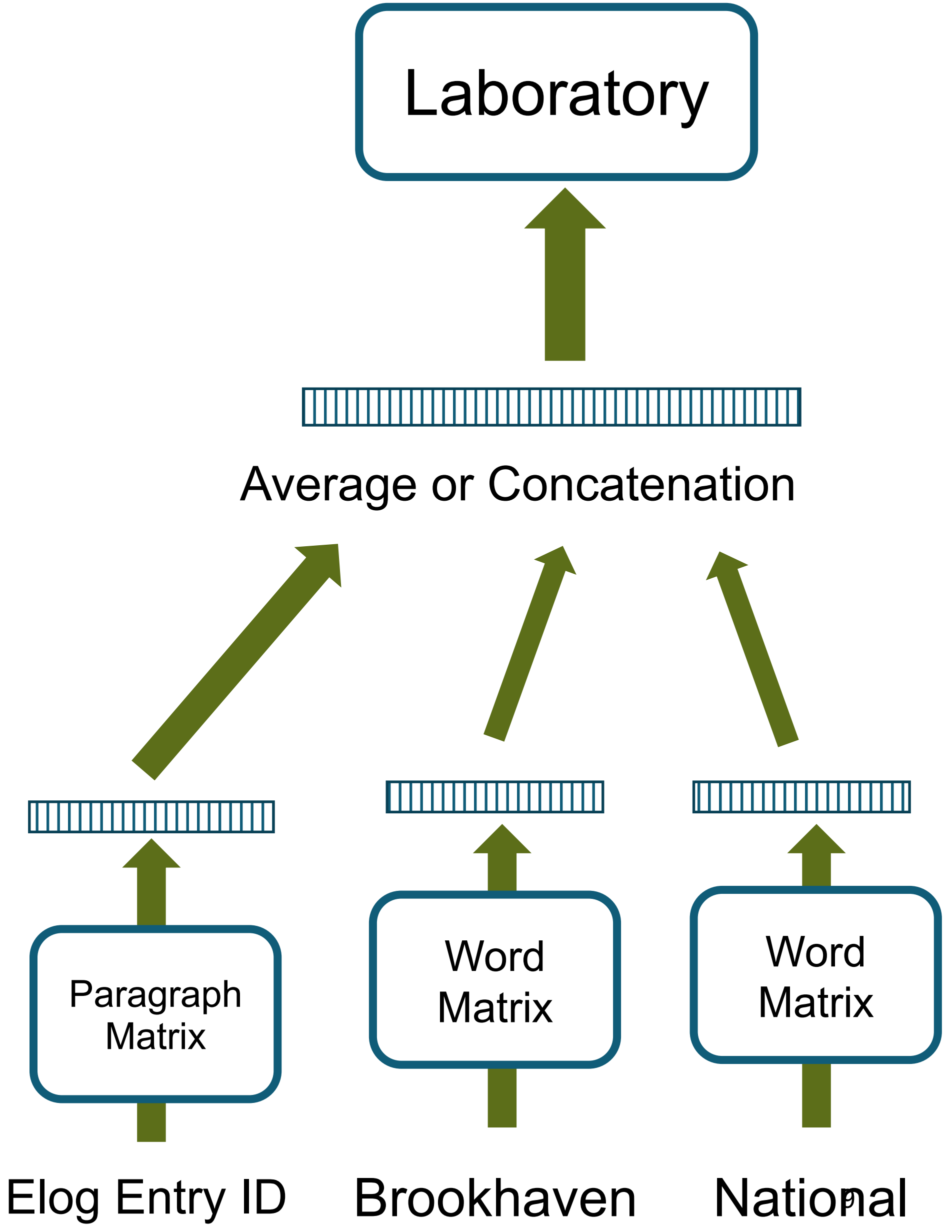
```
Service, building, and equipment tour complete.
```

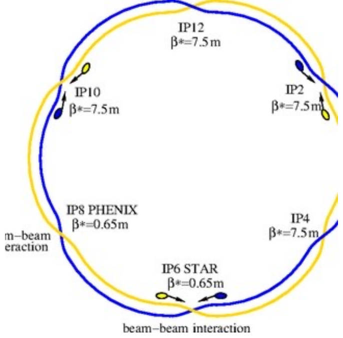
```
[service, building, equipment, tour, complete]
```

Models

Doc2Vec

- Gensim D2V model based on *Distributed Representations of Sentences and Documents* by Le and Mikolov
- Paragraph vectors predict the next word given a sample of words from the paragraph
- Every paragraph is mapped to unique vector which is a column in the paragraph matrix
- Every word in each unique vector also gets mapped to a unique column vector in the word matrix
- Take processed entries and create a list of d2v tagged documents



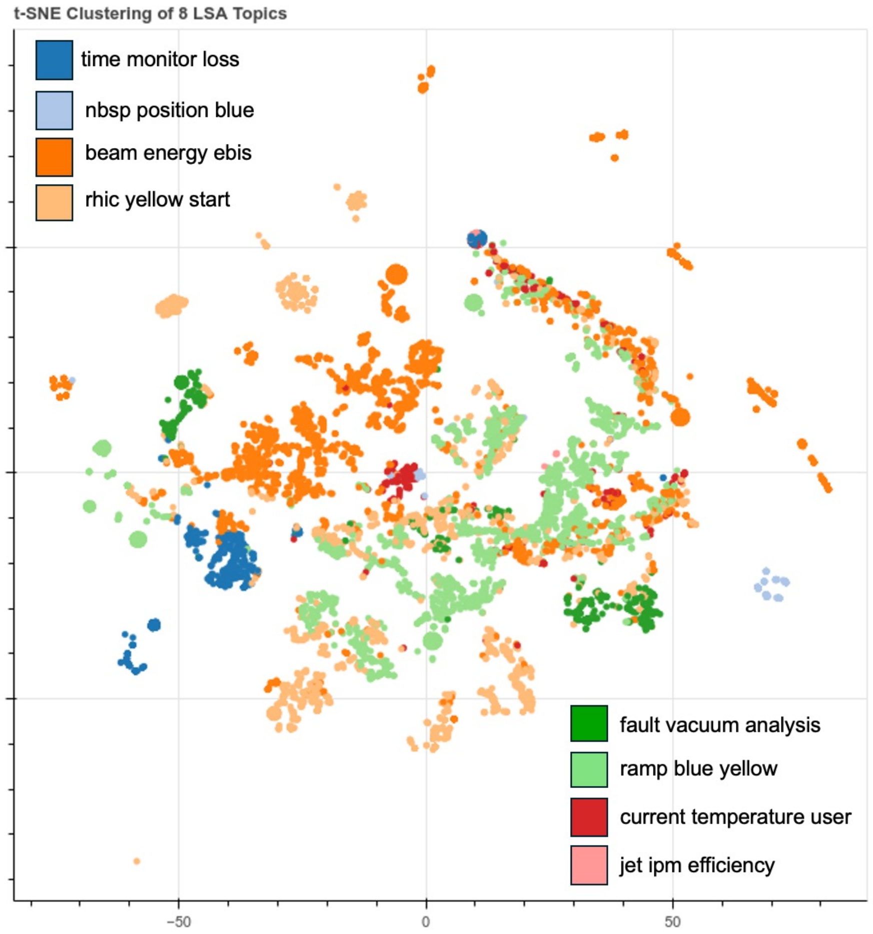


Similar Documents

`polarization for yellow 2h target1 store energy before physics declared yellow beam intensity`

1. Yellow 1 V6 Polarization: -51.53 6.05% 78%
   Yellow 2 H6 Polarization: -51.28 10.83%
2. Polarization For Yellow 1 V Target2: 51.44 &plus mn 1.94 Store Energy (254.21) Before Physics Declared, Yellow Beam Intensity: 208.3x10^11 77%
3. Yellow 1 V5 Polarization: -56.67 4.87% Yellow 2 H5 Polarization: -59.46 6.09% 76%

Topic Modeling



Latent Semantic Analysis

LSA uses dimension reduction techniques to find meanings and similarities of documents by how frequently words appears in those documents.

Latent Dirichlet Analysis

LDA utilizes vector representations of the ratio of the counts of words in document data.

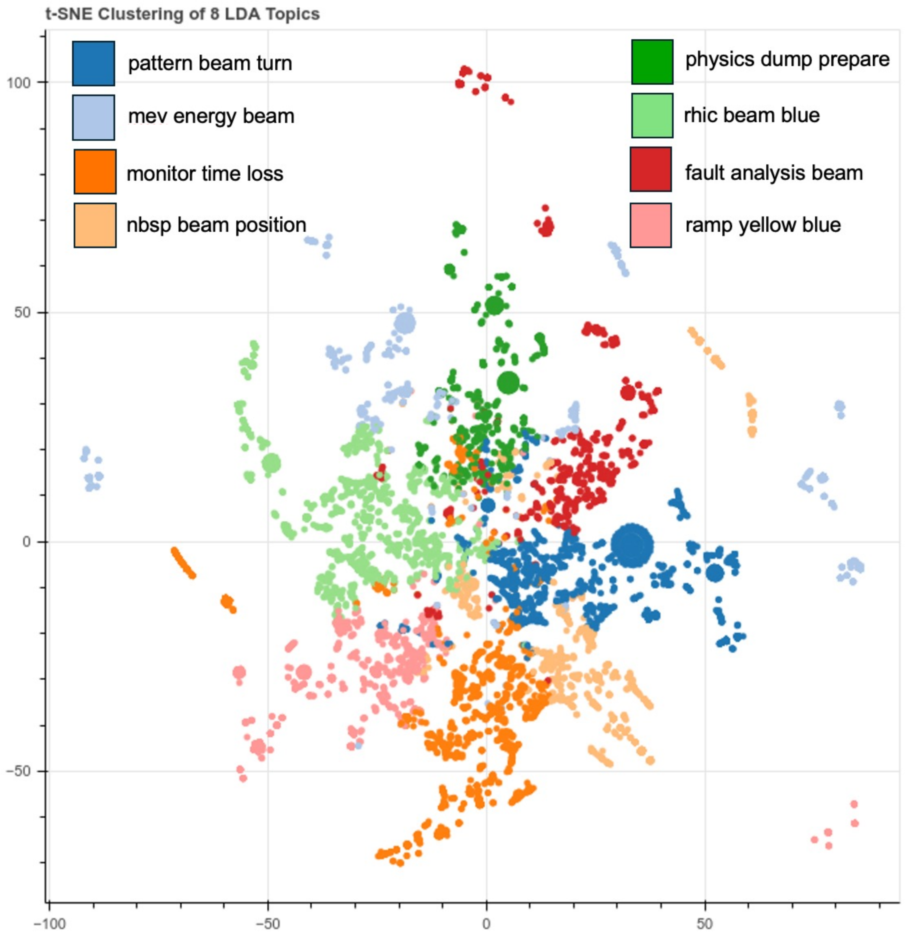
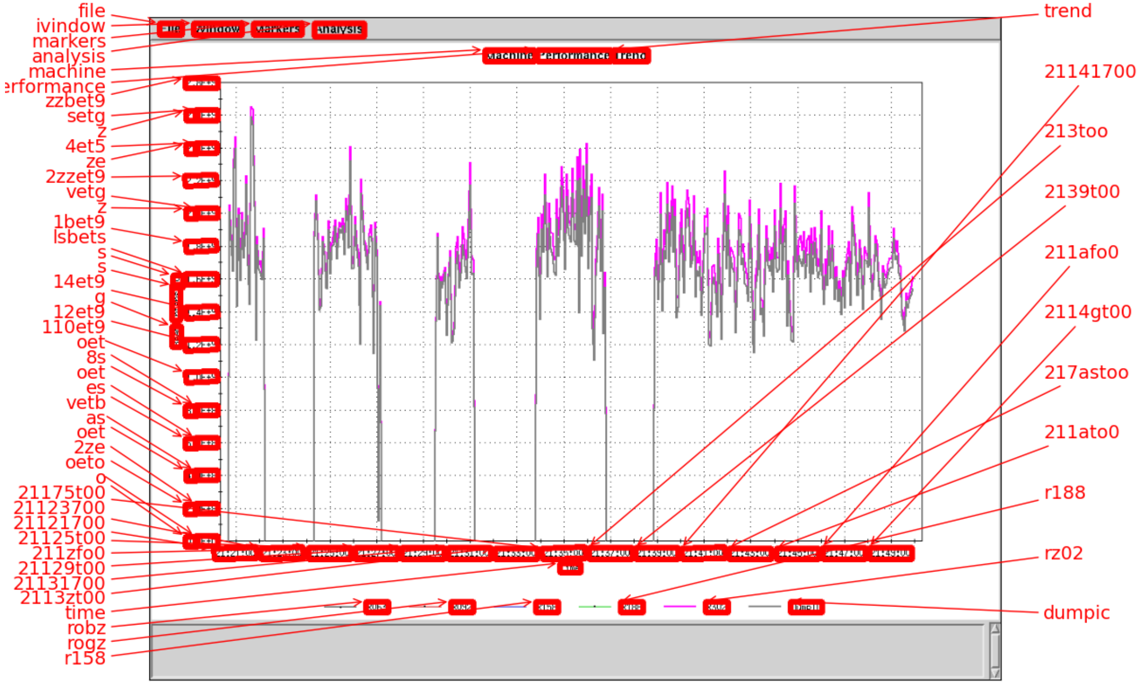
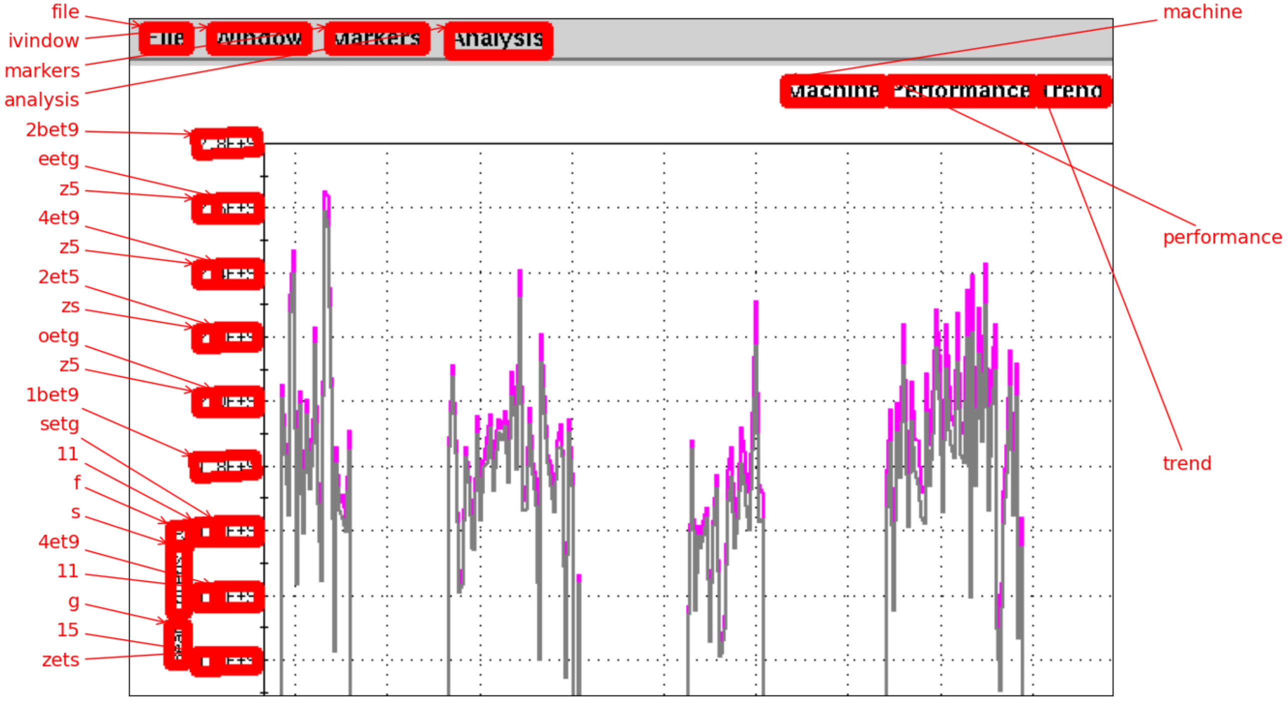


Image Processing

Optical character recognition (OCR) is utilized to parse text out of images to be used for natural language processing tasks or many other applications. Keras OCR was used to test text recognition on images attached to elog entries.

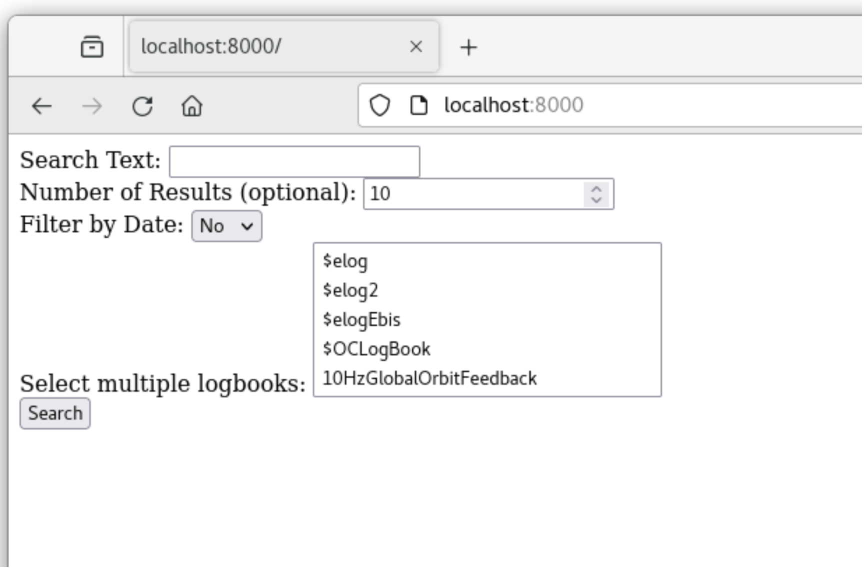




User Interface

User Interface

- Wanted a web demo for users to test and give feedback
- Created a simple website using FastApi and Jinja templates

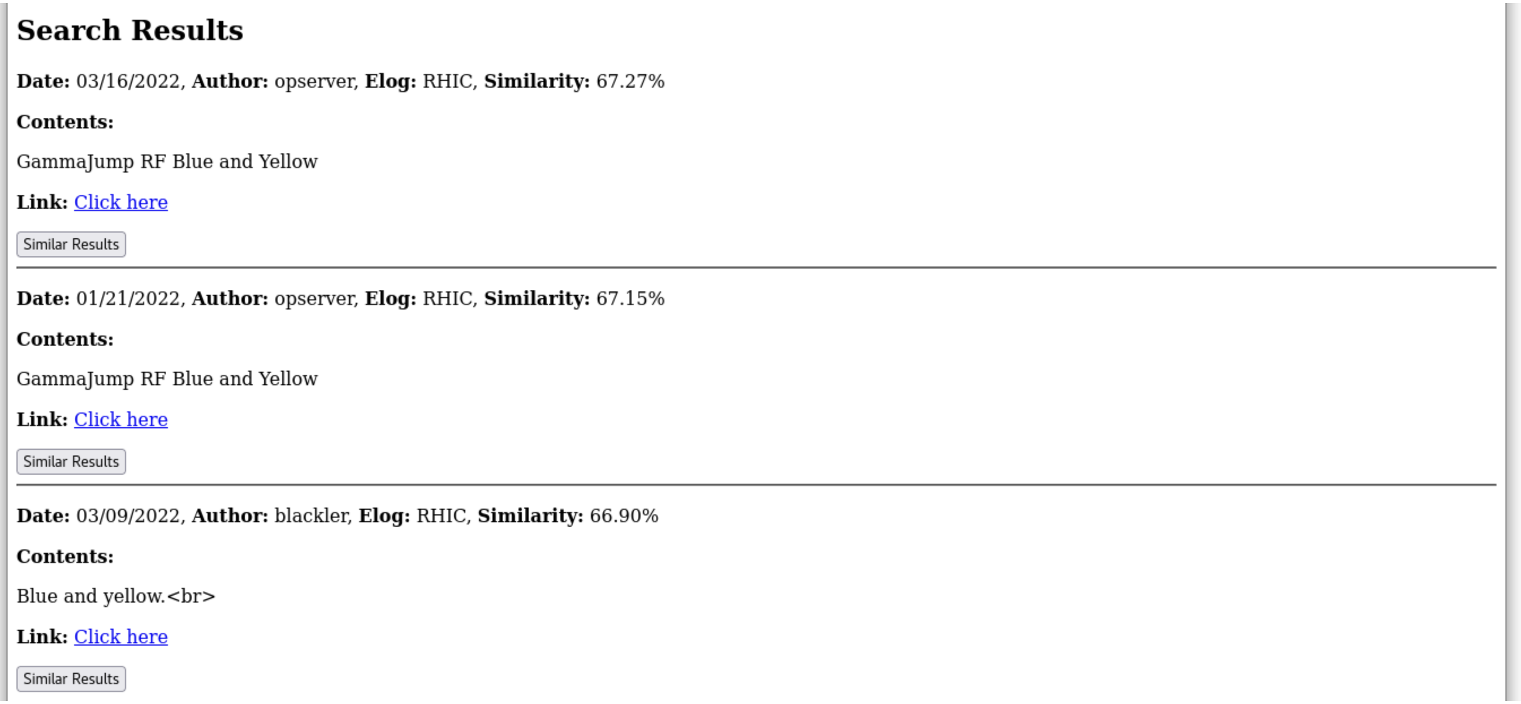


Search Text:

Number of Results (optional): 10

Filter by Date: No

Select multiple logbooks:
$elog
$elog2
$elogEbis
$OCLogBook
10HzGlobalOrbitFeedback



Search Results

Date: 03/16/2022, **Author:** opserver, **Elog:** RHIC, **Similarity:** 67.27%

Contents:

GammaJump RF Blue and Yellow

Link: Click here

Date: 01/21/2022, **Author:** opserver, **Elog:** RHIC, **Similarity:** 67.15%

Contents:

GammaJump RF Blue and Yellow

Link: Click here

Date: 03/09/2022, **Author:** blackler, **Elog:** RHIC, **Similarity:** 66.90%

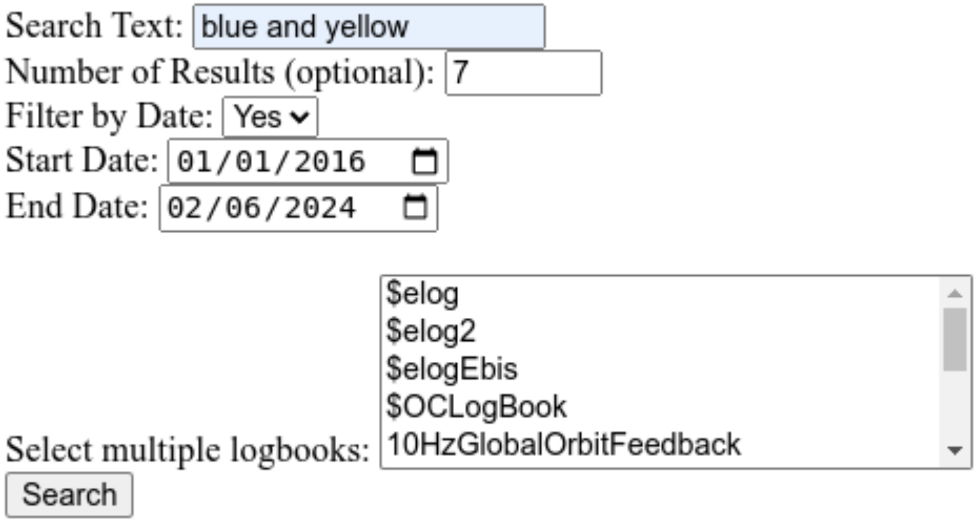
Contents:

Blue and yellow.

Link: Click here

User Interface

- User's can search by elog and date
- Filter number of responses
- This page will be linked to the elog for initial testing
- The current elog system will be evaluated for EIC to see if this functionality will be added to the elog or another tool



Search Text: blue and yellow

Number of Results (optional): 7

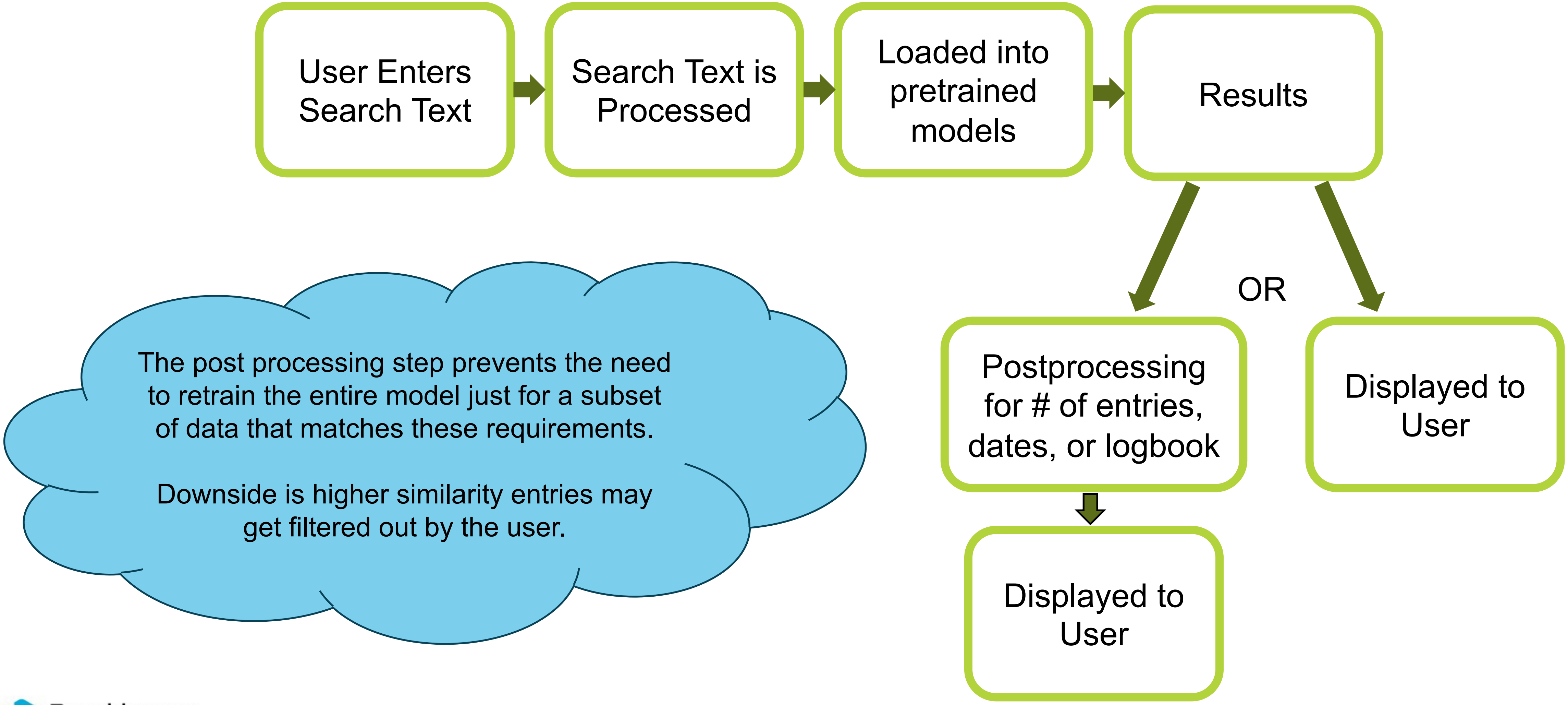
Filter by Date: Yes

Start Date: 01/01/2016

End Date: 02/06/2024

Select multiple logbooks:
$elog
$elog2
$elogEbis
$OCLogBook
10HzGlobalOrbitFeedback

Web Interface Workflow



User Enters Search Text

Search Text is Processed

Loaded into pretrained models

Results

OR

Postprocessing for # of entries, dates, or logbook

Displayed to User

Displayed to User

The post processing step prevents the need to retrain the entire model just for a subset of data that matches these requirements.

Downside is higher similarity entries may get filtered out by the user.

What’s Next?

- Link this demo webpage into the elog system to allow users to test thoroughly
    - Feedback drives new features
- Then decide how to implement directly into the elog (or another tool)
- Student to development reinforcement learning model
    - Improve searches with user interaction!

Thank you!

Thanks to Sam Clark, Wenge Fu, and Seth Nemesure for their knowledge and support.