Enhancing Electronic Logbooks using Machine Learning

Jennefer Maldonado, Samuel Clark, Wenge Fu, Seth Nemesure
Collider Accelerator Department, Brookhaven National Laboratory, Upton, New York, United States

Abstract

The electronic logbook (elog) system used at Brookhaven National Laboratory's Collider-Accelerator Department (C-AD) allows users to customize logbook settings, including specification of favorite logbooks. Using machine learning techniques, customizations can be further personalized to provide users with a view of entries that match their specific interests. We will utilize natural language processing (NLP), optical character recognition (OCR), and topic models to augment the elog system. NLP techniques will be used to process and classify text entries. To analyze entries including images with text, such as screenshots of controls system applications, we will apply OCR. Topic models will generate entry recommendations that will be compared to previously tested language processing models. We will develop a command line interface tool to ease automation of NLP tasks in the controls system and create a web interface to test entry recommendations. This technique will create recommendations for each user, providing custom sets of entries and possibly eliminate the need for manual searching.

Data Preprocessing

	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

Tokenize

Split each elog entry into individual words

```
This is an elog entry.
```

```
[this, is, an, elog, entry]
```

Lemmatize

Group different forms of one word into the same form

```
log logs logging logged
```

```
log
```

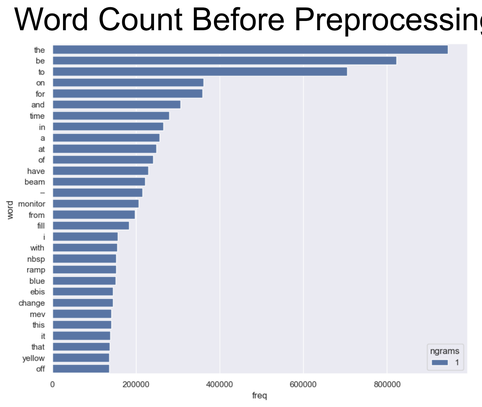
Remove Stop Words

Remove commonly used words in the English language

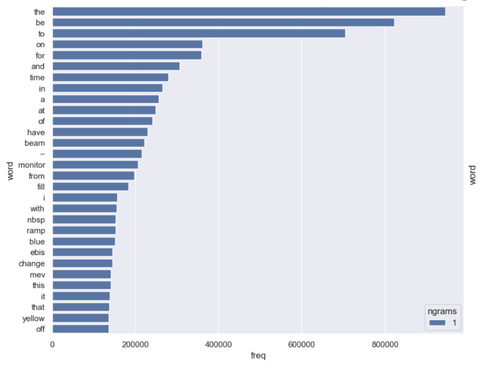
```
This is an elog entry.
```

```
[elog, entry]
```

Word Count Before Preprocessing



Word Count After Preprocessing

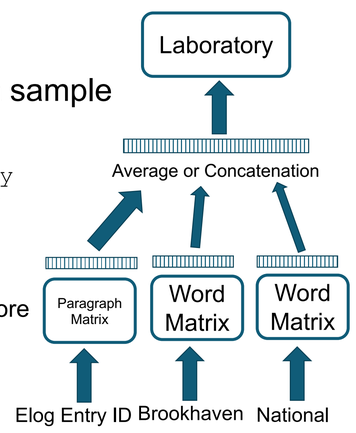


Doc2Vec Model

Paragraph vectors predict the next word given a sample of words from the text.

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

1. Yellow 1 V6 Polarization: -51.53 6.05% 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
3. Yellow 1 V5 Polarization: -56.67 4.87% Yellow 2 H5 Polarization: -59.46 6.09%



Classification

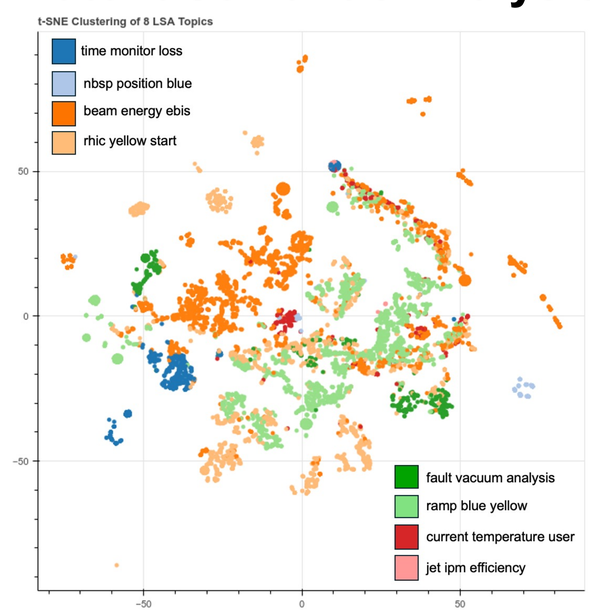
Multinomial Naïve Bayes for multinomially distributed data.
Accuracy Score: ~78% Recall Score: ~74%
Precision Score: ~66% F-Score: ~0.70

R-Failure
Alarm cleared by access control personnel. Might have been related to power lost or the fire alarm testing.

F-MachineSetup
Received call to inform us that the work at the Booster argon station #1 is completed. It is now back to normal operations.

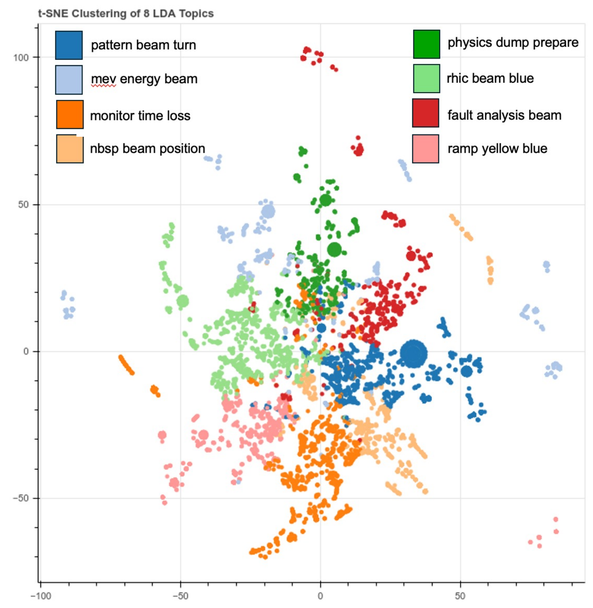
Topic Modeling

Latent Semantic Analysis



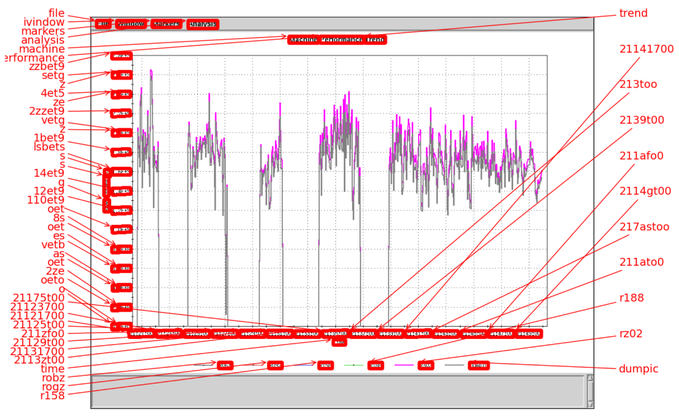
LSA uses dimension reduction techniques to find meanings and similarities of documents by how frequently words appears in those documents.

Latent Dirichlet Allocation



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

Image Processing

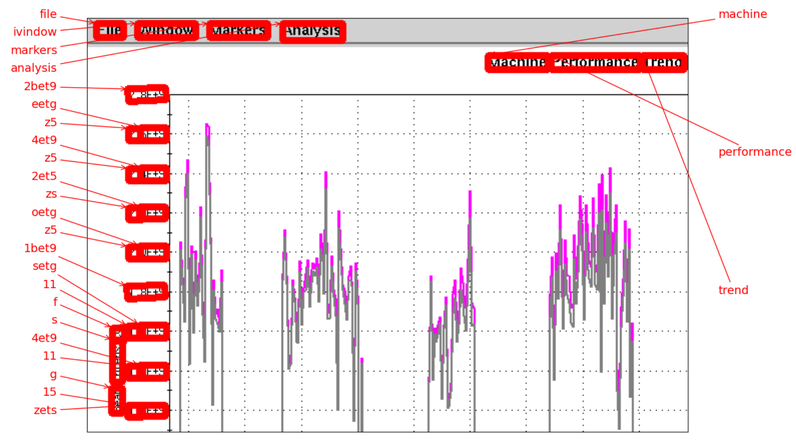


Optical character recognition (OCR) is utilized to parse text out of images to be used for natural language processing tasks or many other applications. Keras OCR was used to test text recognition on images attached to elog entries.

Left: machine performance trend graph upload to an elog entry after keras-ocr analysis
Below: close up of graph and analysis

Ease of Use:

```
# load images
keras_ocr.tools.read("image.png"
)
# load pretrained model
pipeline=keras_ocr.pipeline.Pipe
line()
# get predictions
preds =
pipeline.recognize(images)
# use the predictions to draw
labels
keras_ocr.tools.drawAnnotations(
images, preds)
```



Conclusion

Natural language processing makes searching large databases of text much easier. The electronic logbook system is a good example of how these methods improve user's experience and search results. The future goals of this project are to release a web-based search engine where users can enter a word or phrase. The text will be processed, and similar entries will be produced by the Doc2Vec model and classifier. All the techniques discussed have improved the organization of the data and will benefit user's experiences using the elog system.

References

1. I. Jordan, M. Blei, A. Ng. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022, 2003.
2. F. Chollet et al. Keras-OCR, 2015.
3. S. Dumais. Latent Semantic Analysis. Annual Review of Information Science and Technology, 38(1):188-230, 2004
4. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J.Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12:2825-2830, 2011.
5. R. Rehurek and P.Sojka. Gensim: Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* pages 45-50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Contact

Jennefer Maldonado, jmaldonad@bnl.gov
Senior Applications Analyst, Collider Accelerator Department

U.S. DEPARTMENT OF ENERGY

www.bnl.gov

Brookhaven National Laboratory