Japanese Text Analysis Using Splunk

3.1 Introduction

We adopted Splunk[*1] on IIJ xSP Platform Service/Mail, a large-scale email service with millions of accounts, to extract useful information from the huge amount of logs generated, perform systems analysis, and to protect users from spammers.

We initially used it mainly for searching logs, but our wide uses for Splunk now include automated spam detection using the Splunk Machine Learning Toolkit (Figure 1)[*2] and the streamlining of services operations.

Below, we start by giving some background to our adoption of Splunk. We then go over the Japanese language processing extension we built for NLP in the Splunk Deep Learning Toolkit, which Splunk merged into the toolkit, and describe text mining using it.

3.2 Background to Adopting Splunk

On the IIJ xSP Platform Service/Mail service, customer support center staff perform log analyses in response to end-user requests, including email delivery searches, the display of individual email delivery routes, and Web mail and POP/IMAP/SMTP authentication log searches. The functionality to do this was implemented in ElasticSearch. We were also using ElasticSearch and Kibana for other service operating tools within IIJ. Back when the service was launched, for instance, we used them to identify users generating large levels of input, detect errors, and create reports for customers.

Partly because we sensed certain limitations with ElasticSearch and Kibana, we settled on Splunk when looking to introduce machine learning algorithms to improve the accuracy of spam detection in the aim of further improving

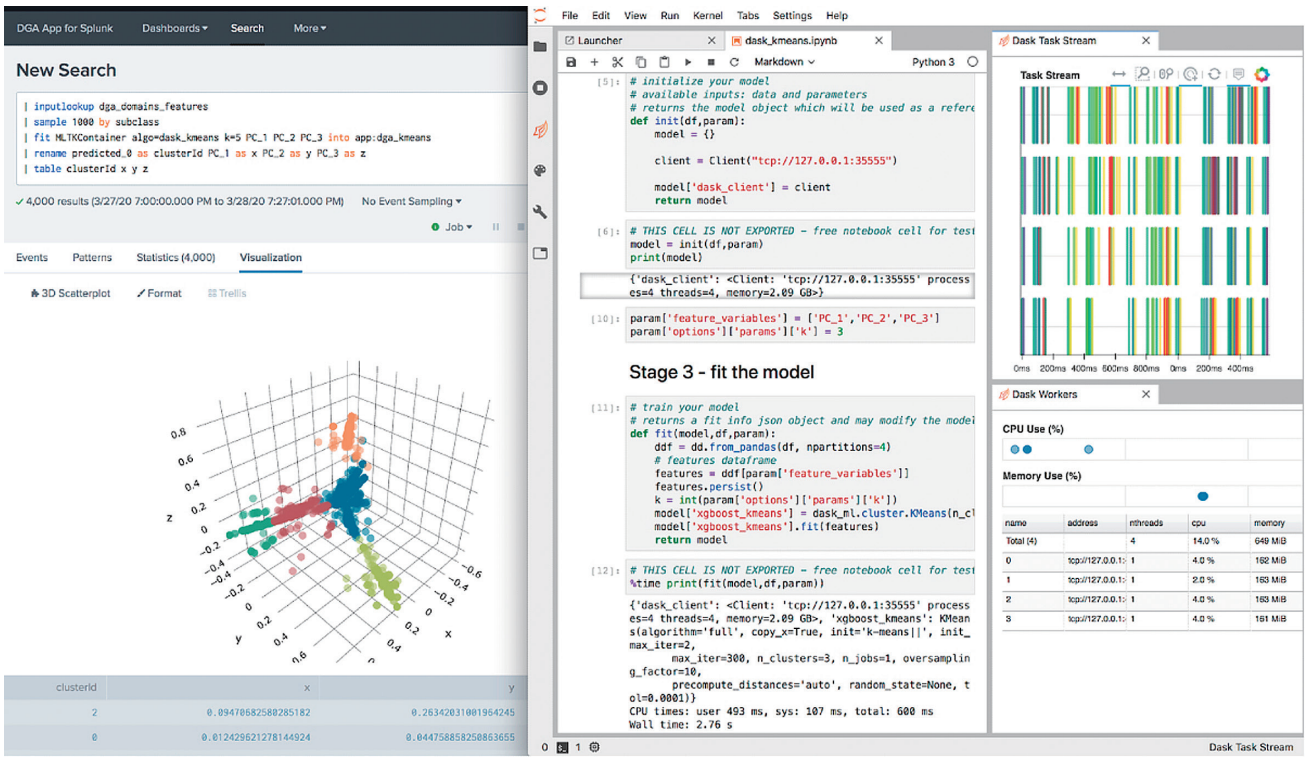


Figure 1: Overview of the Splunk Machine Learning Toolkit

*1 Splunk Enterprise: Find out what is happening in your business and take meaningful action quickly (https://www.splunk.com/en_us/software/splunk-enterprise.html).

*2 Splunk Machine Learning Toolkit (https://www.splunk.com/en_us/blog/machine-learning/deep-learning-toolkit-3-1-examples-for-prophet-graphs-gpus-and-dask.html).

the quality of service on IIJ xSP Platform Service/Mail. Our reasons were that Splunk has a wealth of plugin and visualization apps (both free and paid) optimized for a range of purposes as well as the prospect of speedy development, its offers outstanding system stability and maintainability in comparison with ElasticSearch, and the free Machine Learning Toolkit and Deep Learning Toolkit were appealing.

3.3 Using Splunk for Spam Detection

To improve accuracy using machine learning, in addition to selecting an algorithm, you also need to select axes for analysis, adjust the algorithm parameters, run the learning process, and repeatedly test the model. The Splunk Machine Learning Toolkit and Deep Learning Toolkit provide a user interface that lets you do this seamlessly, and we were able to evaluate algorithms and improve model accuracy in a short period of time.

Spam uses a variety of techniques to blend in among legitimate users. And because activity attributes differ depending on the spam, you need to take an overall view when detecting it (Figure 2).

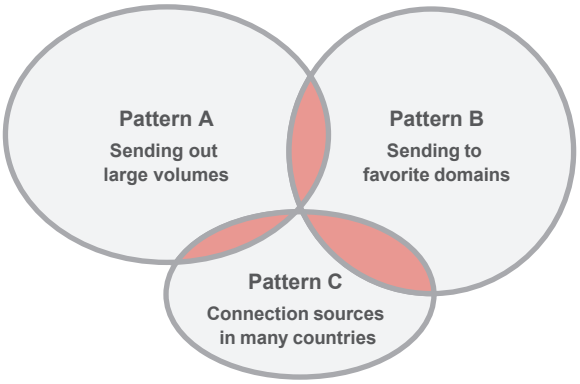
On IIJ xSP Platform Service/Mail, we evaluated a number of algorithms and combinations of variables, including number of source IPs, number of source countries, number of emails sent within a certain time frame, number of unique destinations, whether emails are being sent mainly to domains that spammers like to target or whether emails are also being sent to other domains in a similar manner, and frequency of sending errors. We obtained good results with SVMs[*3]. SVMs are supervised learning models that exhibit robust prediction performance and allow the use of n-dimensional hyperplanes. They also use margin maximization to find the boundary that represents the largest separation from each of the classes being considered.

3.4 The Need for Japanese Text Analysis and NLP (Natural Language Processing)

We have been working to create value-added by analyzing the various logs generated by our services to obtain data useful in the operation and running of those services. Beyond the feature analysis of spam sampled from specific points, we have also heard from other teams here at IIJ that they are, for example, having trouble dealing with abuse or reading in Redmine tickets for analysis, so evidently there is also a need to analyze Japanese text data within IIJ itself.

Applying NLP to text data from abuse emails and Redmine tickets and performing analysis along axes such as “people” and “equipment” makes possible the early discovery of, for example, where loads and problems are concentrated.

Splunk can do morphological analysis using MeCab, but processing large amounts of text data and performing advanced



Example: The actual spammers are those that match several patterns, as indicated by the colored-in regions in the figure

Figure 2: Spammer Activity

*3 SVM: Support Vector Machine, a type of machine learning algorithm.

text mining with this alone is difficult. So we thought about using NLP from the Splunk Deep Learning Toolkit. Using NLP makes it possible to do sentence structure analysis, extract named entities, and so forth, and we were heavily drawn to the prospect of being able to read in large amounts of text data for mining. Named entity extraction is a technique that seeks to identify named entities (objects that can be denoted with a proper name) in text and classify them according to predefined attributes into categories (entity types) such as people, organizations, place names, dates, and numbers (Figure 3).

At the time we started testing, the Splunk Deep Learning Toolkit did not support Japanese NLP, so we had to create our own extension for Japanese, which we published on Splunkbase, Splunk's official library. The extension has now been merged into the Splunk Deep Learning Toolkit. Adding Japanese NLP support to the Splunk Deep Learning Toolkit and thereby broadening the scope for its use in business in Japan generated a considerable response from people. I had the opportunity to give a presentation at a GOJAS (Go Japan Splunk User Group) event to an audience of over 100 (Figure 4).

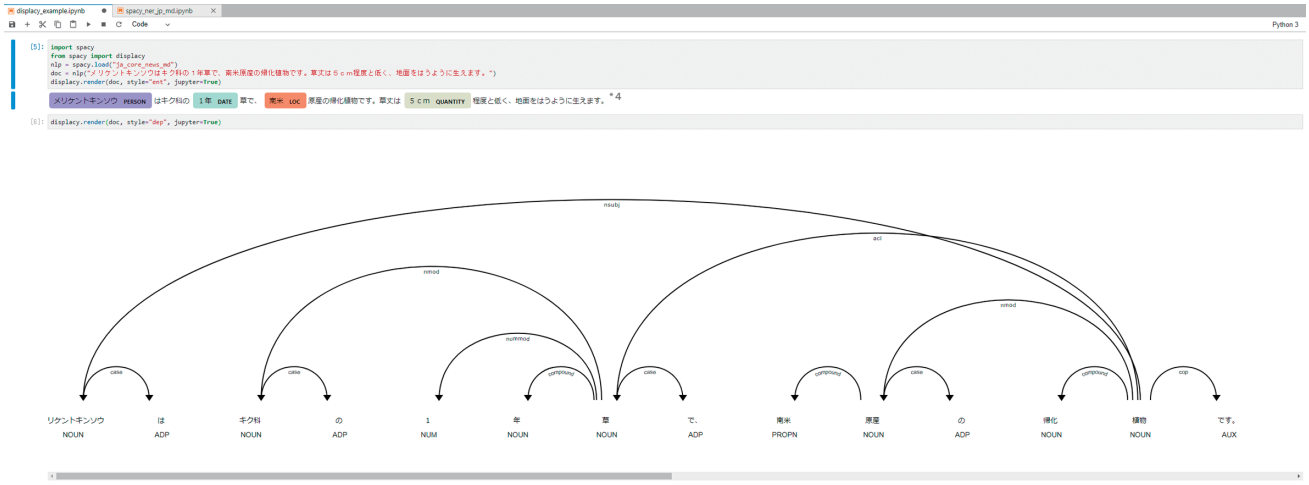
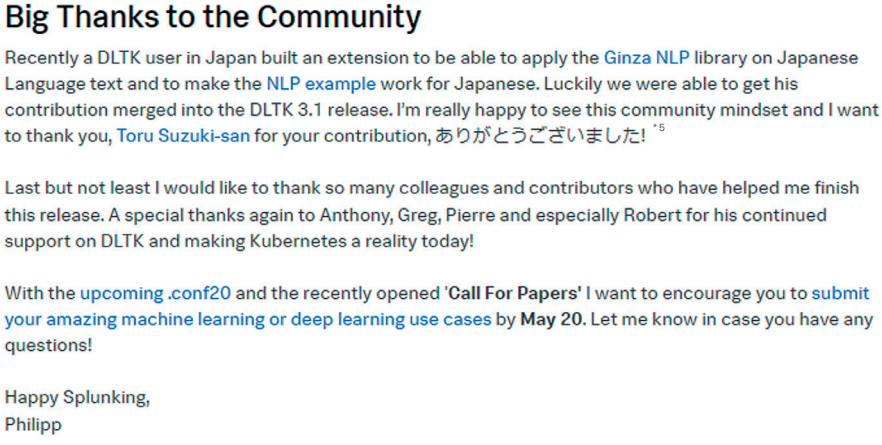


Figure 3: Example of Named Entity Extraction in Jupyter



Big Thanks to the Community

Recently a DLTK user in Japan built an extension to be able to apply the Ginza NLP library on Japanese Language text and to make the NLP example work for Japanese. Luckily we were able to get his contribution merged into the DLTK 3.1 release. I'm really happy to see this community mindset and I want to thank you, Toru Suzuki-san for your contribution, ありがとうございました! *5

Last but not least I would like to thank so many colleagues and contributors who have helped me finish this release. A special thanks again to Anthony, Greg, Pierre and especially Robert for his continued support on DLTK and making Kubernetes a reality today!

With the upcoming .conf20 and the recently opened **'Call For Papers'** I want to encourage you to submit your amazing machine learning or deep learning use cases by **May 20.** Let me know in case you have any questions!

Happy Splunking,
Philipp

Figure 4: Message from the Splunk Deep Learning Toolkit on Our Contribution*6

*4 For English-speaking readers, the output of the displacy.render function is translated from the original Japanese.

*5 The Japanese text in the message reads: Thank you very much!

*6 splunk.com, "Deep Learning Toolkit 3.1 - Examples for Prophet, Graphs, GPUs and DASK" (https://www.splunk.com/en_us/blog/machine-learning/deep-learning-toolkit-3-1-examples-for-prophet-graphs-gpus-and-dask.html).

3.5 Text Mining with NLP (Natural Language Processing)

Text mining with NLP involves getting an overall picture of a passage of text and performing feature extraction based on information obtained by analyzing the relationships between words and extracting named entities.

NLP in the Splunk Deep Learning Toolkit works in conjunction with Jupyter running in a Docker container, with the algorithms implemented in spaCy, a Python NLP library.

Entity	Entity_Count	Entity_Type	Entity_Type_Count
183万円	150	MONEY	42
1億円	96	MONEY	42
5月5日	96	DATE	55
92%	95	QUANTITY	108
日本	87	GPE	15
1万円	63	MONEY	42
9割	63	PERCENT	20
100%	56	QUANTITY	108
250万円	54	MONEY	42
100万円	52	MONEY	42
15分	49	TIME	16
1つ	43	QUANTITY	108
100人	42	QUANTITY	108
4000万円	41	MONEY	42
10分間	36	TIME	16
100%	34	PERCENT	20
火	33	DATE	55
11年	32	DATE	55
30万人	32	MONEY	42
第2267号	32	ORDINAL	10
800人	31	QUANTITY	108
橋本純樹	31	PERSON	45
3000万円	30	MONEY	42
92%	29	PERCENT	20
ワンクリックスキャル24/7完全無料公開中	28	PRODUCT	19
1割	25	PERCENT	20

Table 1: Named Entity Recognition Results for Spam Sampled at a Fixed Point on May 1, 2020 (Entity column translated from the original Japanese)

To enable the processing of Japanese text, we customize the Docker container image, upgrading to spaCy 2.3.2 and installing language models with the Japanese models added in.

The named entity recognition algorithm is written in a Jupyter notebook as so is easily customizable.

Table 1 shows the results of analyzing text data from a single day (May 1, 2020) of spam sampled at a set point using the named entity recognition algorithm we extended. We use the ja_core_news_md model (see https://spacy.io/models/ja for details). Entity denotes a named entity, Entity_Count is the number of times the named entity appears, Entity_Type is a collection of entities having similar attributes defined in the model, and Entity_Type_Count is the number of occurrences of that Entity_type.

The analysis extracts entities such as people (PERSON), monetary amounts (MONEY), place names (GPE), dates (DATE), times (TIME), and quantities (QUANTITY). It is worth noting that strings representing PRODUCT entities are extracted without being broken into separate words.

The table is sorted in descending order of Entity_Count, and the Entity_Type column shows that MONEY entities occupy top spots in the list, indicating that a lot of the spam on this day contained content relating to monetary amounts.

The names are extracted without being split into first name and surname, which is a great advantage when performing analysis along the personal name axis. Since named entity recognition lets us classify large amounts of text data by personal names or product names, it could, for example, be used to analyze operating status or turn text-based knowledge into a database.

Next, to see what differences appear between samples of spam taken in February and May 2020 at a fixed point, we graph the top 15 named entity recognition from those samples. Figures 5 and 6 show the results.

The entity English:LANGUAGE ranked at the top in February and was a clear standout in relative percentage terms as well, perhaps reflecting that overseas travel was still happening during the early stages of the COVID-19 situation. In May, once Japan had declared a state of emergency, English:LANGUAGE had fallen heavily in the ranking, being replaced by MONEY entities, which had also increased substantially in absolute terms, indicating a rise in spam activity.

The analysis of text data is hindered by the lack of classifying information and difficulty nailing down analysis axes, but using named entity recognition like this lets us classify text data using the attributes of named entities, and this makes it a highly valuable technique.

And using the combination of named entity and attribute classification makes it possible to identify overall patterns in text, which greatly opens up the possibilities for text mining.

Entity	Entity_Count	Entity_Type
橋本純樹	31	PERSON
佐々木千恵	22	PERSON
エリオット	17	PERSON
プロスペクト	17	PERSON
橋本	17	PERSON
佐々木	15	PERSON
トニー野中	9	PERSON
北条	9	PERSON
良彰	9	PERSON
アダム	8	PERSON
ロスチャイルド	8	PERSON
倉持	8	PERSON
サトー	7	PERSON
木村	7	PERSON
村岡	7	PERSON
よしあき	5	PERSON
ベール	5	PERSON
ザラ	4	PERSON
スキャルロジック	3	PERSON
たかはしよしあき	2	PERSON
カリスマ美人	1	PERSON
友宮真	1	PERSON
堀崎むつみ	1	PERSON
塚弥生	1	PERSON
大元大輝	1	PERSON

Table 2: PERSON Entities Found Using Named Entity Recognition on Spam Sampled at a Fixed Point on May 1, 2020 (Entity column translated from the original Japanese)

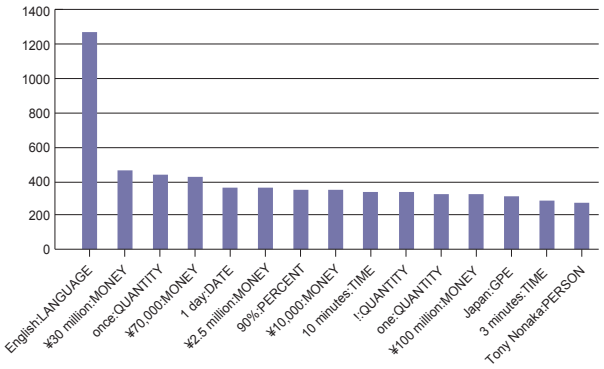


Figure 5: Graph of Top 15 Named Entity Recognition Results for February 2020

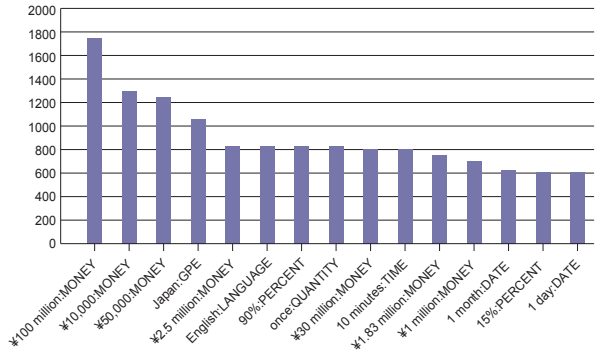


Figure 6: Graph of Top 15 Named Entity Recognition Results for May 2020

3.6 Business Use cases of NLP and Text Mining

Text mining is generally used to discover potential/latent needs based on data accumulated from various text data sources.

Voice data can also be converted to a text data source using external voice-to-text APIs, so voice data accumulated from call center operations and the like can also be used in, for example, customer insight analysis and knowledge extraction for business operations. There exist use cases that involve building a database of examples from text data and matching them by searching for similar patterns, and these approaches are used not only in needs discovery but also in applications such as performance evaluations based on content similarity.

Other companies use text mining in their service operations, an example being the use of a chatbot to serve as the primary contact in a text chat or voice chat, with the interaction being escalated to a human service representative if necessary based on an analysis of the text data generated from the chat. This service approach is used successfully in call center operations, for example, as a labor-saving measure intended to reduce costs.

3.7 Conclusion

In the past, the difficulty in making use of large amounts of text data relegated it to dark data status, but advances in the accuracy of natural language processing have now opened up a wide range of uses for text mining that make it possible to discover useful information.

There are also tools like the Splunk Deep Learning Toolkit that provide a seamless interface for performing natural language processing on text from accumulated data and doing everything from generating models through to text mining. With text mining in the spotlight of late, perhaps now is the time to start using it in your business.

Toru Suzuki
Senior Engineer, Service System Development for xSP Operations Section, Application Service Department, Network Cloud Division, IIJ.
Mr. Suzuki is an administrative member of GOJAS (Go Japan Splunk User Group).
He is engaged in efforts to use Splunk to generate value-added in services.