Splunkによる日本語文章解析処理

3.1 はじめに

数百万アカウントを収容する大規模メールサービスとなるIIJ xSPプラットフォームサービス/Mailでは、大量蓄積するログからの有用な情報抽出・システム解析・迷惑メール送信者と戦うためにSplunk[*1]を導入しました。

導入当初はログ検索を中心に利用していましたが、昨今はSplunk Machine Learning Toolkit（図-1）[*2]を用いたスパム検知自動化、サービス運用の効率化など、幅広くSplunkを活用しています。

今回はSplunkの導入経緯から始まり、Splunk Deep Learning ToolkitのNLPに日本語処理機能を追加拡張しSplunk社にフィードバック・マージされたお話と、これを用いたテキストマイングについて紹介します。

3.2 Splunk導入経緯

IIJ xSPプラットフォームサービス/Mailでは、顧客サポートセンター向けの機能としてメール配送検索、個々のメールの配送経路表示、WEBメール、POP/IMAP/SMTP認証ログの検索機能など、サポート窓口のスタッフがエンドユーザからの問い合わせに対してログを調査する機能を、ElasticSearchを用いて実装しています。またこの他にIIJ社内のサービス運用ツールとして、サービス立ち上げ当初は大量打ち込みを行っているユーザの特定、エラー検出、顧客向けレポート作成などにElasticSearch、Kibanaを活用していました。

IIJ xSPプラットフォームサービス/Mailでは、更なるサービス品質向上を目的に、スパム検知精度を上げるためMachine Languageアルゴリズムの導入検討を進めた際、ElasticSearch、Kibanaに限界を感じていることもあり、

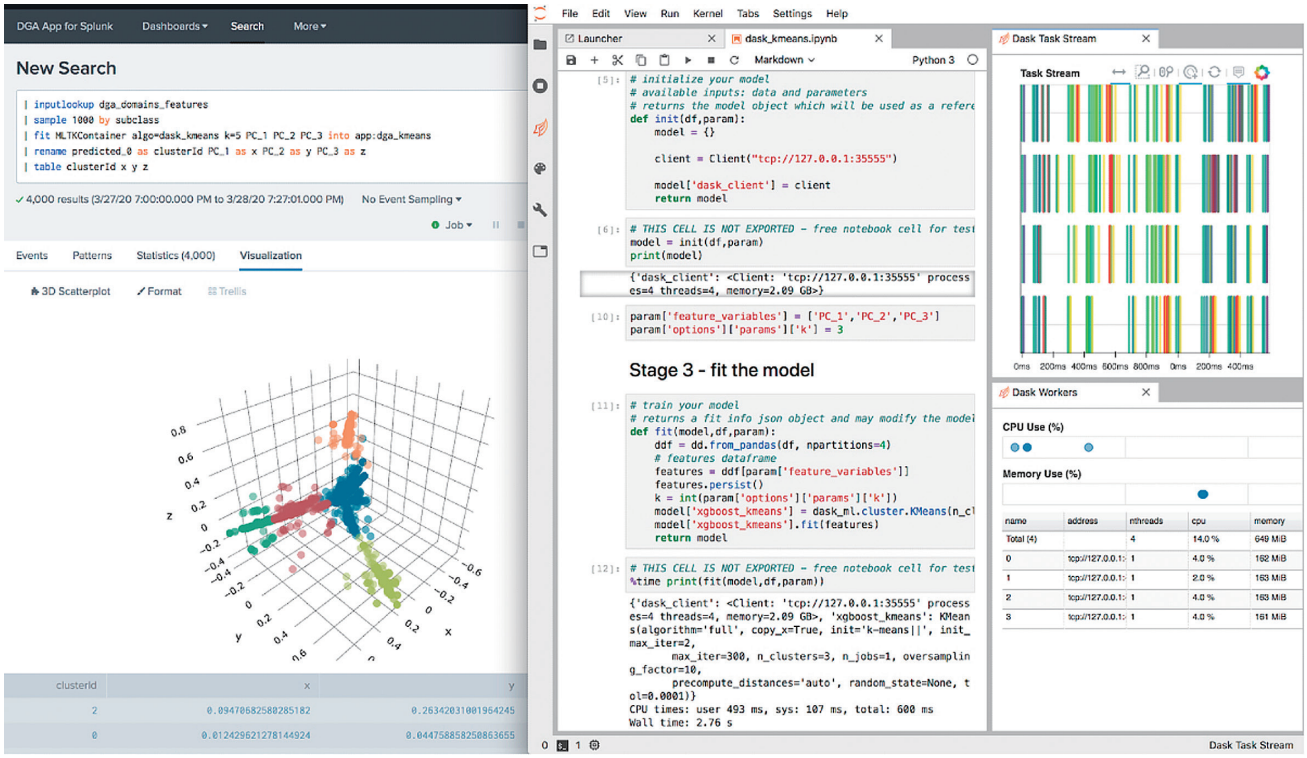


図-1 Splunk Machine Language Toolkitイメージ

＊1　Splunk Enterprise：統合ログ解析・管理ツールビッグデータ分析ソフトウェア（https://www.splunk.com/ja_jp/software/splunk-enterprise.html）。
＊2　Splunk Machine Language Toolkit（https://www.splunk.com/en_us/blog/machine-learning/deep-learning-toolkit-3-1-examples-for-prophet-graphs-gpus-and-dask.html）。

Splunkを導入するに至りました。Splinkは様々な目的に最適化されたプラグイン、可視化Appが豊富(無償/有償)でスピード感のある開発が期待できる上、ElasticSearchと比較して圧倒的なシステム安定性と保守のしやすさがあり、無償のMachine Learning Toolkit/Deep Learning Toolkitが魅力的であることがその理由です。

3.3 Splunkを活用したスパム検知

Machine Learningを用いて精度を上げるためにはアルゴリズムの選択の他、解析軸の選択、アルゴリズムのパラメータ調整、学習、モデルの検証の繰り返し実行が必要ですが、Splunk Machine Learning Toolkit/Deep Learning Toolkitでは、これらがシームレスに実行できるUI環境が提供されており、短期間でアルゴリズムを評価し、モデル精度を上げることができました。

スパムは様々な手法を使って正規ユーザの中に紛れるように活動しています。またはスパムによりActivityの特徴が異なるため、総合的に見て判別する必要があります(図-2)。

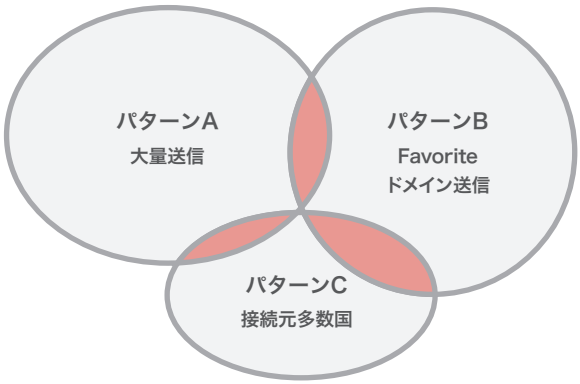
IIJ xSPプラットフォームサービス/Mailでは、送信元IP数、送信元国数、一定時間内における送信数、宛先ユニーク数、スパマが好んでターゲットとするドメインを主に対象として送信しているのか、それ以外のドメインに一様に送信しているのか、送信結果のエラー発生率など、複数変数の組み合わせとアルゴリズム評価を行った結果、SVM[*3]で良い結果を得ることができました。SVMは優れた認識性能を発揮する教師であり学習モデルで、n次元の超平面を扱うことができます。マージン最大化という方法で各クラスから最も遠い境界線を引くという特徴もあります。

3.4 日本語分析ニーズとNLP(Natural Language Processing)

サービスの様々なログを分析することによりサービス運用・運営に有用なデータを得て付加価値創造を目指してきましたが、定点で取得したスパム検体の特徴分析以外にも、ABUSE対応に困っている、Redmineのチケットを取り込んで分析しているなどの声が他部署からあり、社内においても日本語テキストデータを分析するニーズがあることが分かってきました。

ABUSEメールやRedmineチケットのテキストデータをNLPで解析することにより、人や設備などを軸とした分析を行うことで負荷や問題の集中などの早期発見が可能になります。

SplunkでMeCabを使った形態素解析が可能ですが、これだけでは大量のテキストデータの処理や高度なテキストマイニングを行うのは困難です。そこでSplunk Deep Learning ToolkitにあるNLPの利用を考えました。NLPを用いることにより、テキストデータの構文構造解析、固有表現抽出などが可能になり、大量のテキストデータを取り込みテキストマイニングが可能になるところに大きな魅力を感じました。固有表現抽出というのは、テキストから固有表現(Named Entity)を抽出



例:左の色付き部分のように複数パターンの合致部分が真正のスパマ

図-2 スパマのActivityイメージ

*3 SVM:Support Vector Machine。機械学習アルゴリズムの1つ。

し、更に人、組織、地名、日付や数値など、あらかじめ定義されている属性（Entity）に分類、抽出する技術です（図-3）。

検証着手当時Splunk Deep Learning ToolkitのNLPは日本語処理に対応していませんでしたので、独自拡張して日本語対応を行い、SplunkbaseというSplunkの公式ライブラリー上で公開しました。現在はSplunk Deep Learning Toolkitにマージされています。Splunk Deep Learning ToolkitのNLPを日本語対応したことにより、日本でのビジネス活用範囲が広がったことで多くの反響をいただきました。GOJAS（Splunk日本ユーザ会）のイベント講演では、100名を超える方に聴講していただくことができました（図-4）。

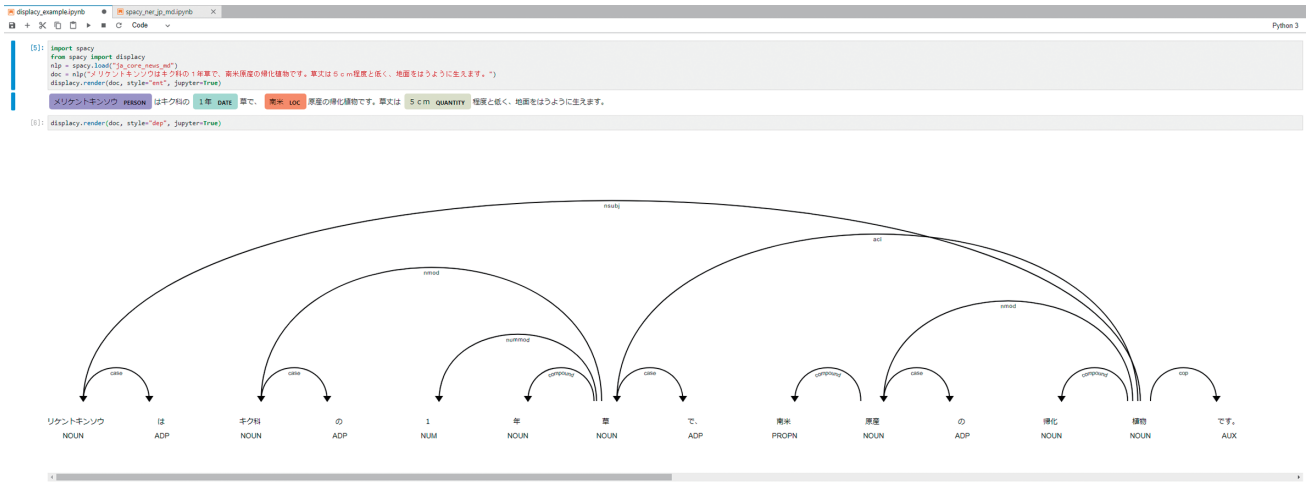


図-3 Jupyter上での固有表現抽出例

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!

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

図-4 寄贈先Splunk Deep Learning Toolkit開発者メッセージ*4

*4 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 NLP(Natural Language Processing)を使ったテキストマイニング

NLPを使ったテキストマイニングでは、語彙間の関係性の分析や固有表現抽出で得られた情報を元に文章の全体像の把握や特徴抽出を行います。

Splunk Deep Learning ToolkitのNLPはDockerコンテナで稼働しているJupyterと連携して動作しており、アルゴリズムはPython Natural Language Processing libraryであるspaCyを用いて実装されています。

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

表-1 2020年5月1日に定点で受信したスパム検体の固有表現抽出結果例

日本語テキストを処理可能にするため、Dockerコンテナイメージをカスタマイズし、spaCy 2.3.2へのアップグレードと追加された日本語モデルを含めた各国言語モデルの導入を行っています。

固有表現抽出のアルゴリズムはJupyter notebookで記述されているため、容易にカスタマイズが可能です。

表-1は定点で受信した2020年5月1日の1日分のスパム検体の本文データを独自拡張した固有表現抽出アルゴリズムで分析した結果になります。モデルはja_core_news_md(詳細はhttps://spacy.io/models/jaを参照)を使用しています。Entityが固有表現、Entity_Countがその固有表現の出現数、Entity_Typeが使用したモデルの中で定義されている属性分類、Entity_Type_Countがその属性分類の出現数を示しています。

人(PERSON)、お金(MONEY)、地名(GPE)、日付(DATE)や時間(TIME)数量(QUANTITY)などが抽出されています。プロダクト(PRODUCT)に該当する文字列が単語に分解されずに抽出されている点が注目されます。

この表ではEntity_Count数が大きい順にソートして出力していますが、Entity_Typeの箇所を見るとMONEYが上位を占めており、この日のスパムはお金に関する内容が記載されているものが多かったことが分かります。

表-2は同一の日のスパムを分析し、人名の属性を示すPERSONで絞り込んだ結果の抜粋です。人名が姓と名に分解されずに抽出されており、人名を解析軸として分析する場合に大きなメリットになります。大量のテキストデータを固有表現抽出により人名やプロダクト名で分類することができるため、稼働状況の分析やナレッジのデータベース化などに活用できそうです。

次に定点取得したスパム2020年の2月分と5月分の固有表現抽出結果でどのような差異が現れるか調べるためにそれぞれ上位15件の固有表現をグラフ化してみると、図-5と図-6のような結果となりました。

2020年2月ではまだコロナ禍の初期で海外旅行も行われていたことを反映しているのか、英語:LANGUAGEが最上位で相対的な比率でも突出して多い状況が分かりますが、緊急事態宣言後の5月では英語:LANGUAGEは大分ランキングを落として属性MONEYのものと入れ替わり、絶対数自体も大分増えていてスパム活動が活発化していることも分かります。

テキストデータ分析が難しい背景に、分類情報がなく解析軸が定まらないという点がありますが、このように固有表現抽出を用いることにより、テキストデータを固有表現の属性を使って分類可能になるので非常に大きな意義があります。

また、固有表現と属性分類の組み合わせ情報を利用することでテキストの概要パターンを識別可能になるため、テキストマイニングの可能性が大きく広がります。

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

表-2 2020年5月1日に定点で受信したスパム検体で固有表現抽出を行い、PERSONで絞り込んだ結果の例

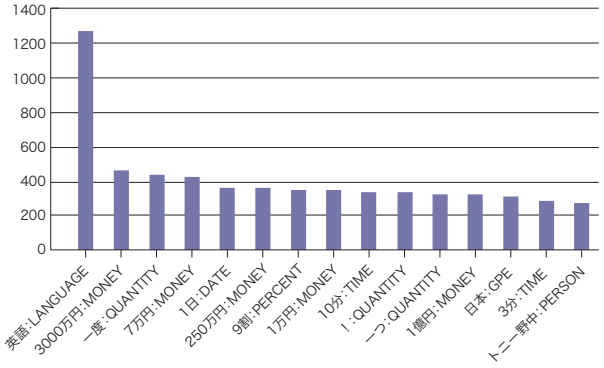


図-5 2020年2月の固有表現抽出結果上位15件のグラフ

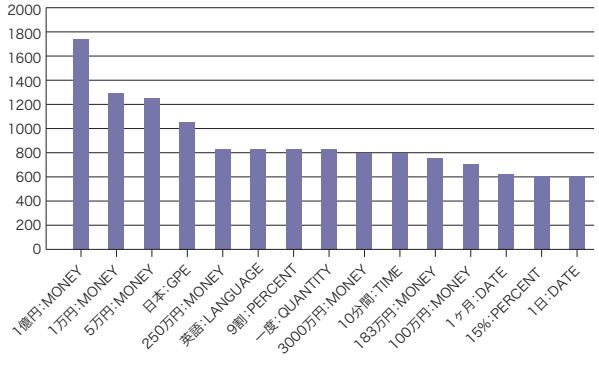


図-6 2020年5月の固有表現抽出結果上位15件のグラフ

3.6 テキストマイニングのビジネス活用

一般的にテキストマイングは様々なテキストデータをソースとして蓄積されるデータを元に、潜在ニーズの掘り起こしを目的として活用されています。

外部の音声テキスト変換APIなどを利用して音声データをテキストデータに変換し、ソースとすることも可能ですので、コールセンター業務などで蓄積される音声データを元にした顧客インサイト分析、業務上のナレッジ抽出などにも活用されています。テキストデータから事例のデータベースを構築し、似通ったパターンの事例を検索してマッチングするなどのユースケースがありますが、これらはニーズの掘り起こしだけでなく一内容の類似性による実績評価などに活用するケースがあります。

他社のサービス事例では、テキストチャット、音声チャットをチャットボットで一次受けを行い、それらのテキストデータを分析して必要に応じて人間による対応にエスカレーションさせるための仕組みの中で活用されています。例えばコールセンター業務の省人化によるコストダウンを目的としたサービスとして上手く建付けが行われている事例が見られます。

3.7 まとめ

従来大量のテキストデータの活用は難しくダークデータと化していましたが、現在では自然言語処理の精度向上により、テキストマイニングを幅広く活用することで有用な情報の掘り起こしが可能になってきています。

Splunk Deep Learning Toolkitのようにデータ蓄積からテキストの自然言語処理、モデル生成から、テキストマイニングまでシームレスに実行できる環境もあります。昨今注目されているテキストマイニングを始めてビジネスへ活用してみては如何でしょうか。

執筆者:
鈴木 徹(すずき とおる)
IIJ ネットワーククラウド本部アプリケーションサービス部xSPシステムサービス課シニアエンジニア。
GOJAS(日本Splunkユーザ会)運営メンバー。Splunkを活用してサービスに付加価値を生み出す活動を行う。