はじめての自然言語処理(NLP: Natural Language Processing)ハンズオン

アマゾン ウェブ サービス ジャパン

シニアエバンジェリスト 亀田 治伸

はじめに:

このハンズオンでは AWS が提供している、自然言語処理関連の 4 つのサービスを用いて、日本語の処理を体験するハンズオンです。以下のサービスを用います。

- ・Amazon Translate: 高速で高品質なニューラル機械翻訳サービス
- ・Amazon Polly: 深層学習技術を使用したテキスト読み上げサービス
- ・Amazon Transcribe:音声をテキストに変換する自動音声認識サービス
- ・Amazon Comprehend: テキストから洞察を見つけるサービス
- ・Amazon Elasticsearch Service: Elasticsearch, Kibana のマネージドサービス
- ・AWS Cloud9: クラウドベースの統合開発環境 (IDE)
- ・AWS Lambda: サーバレスコンピューティングサービス

作業は特に指示されない限り、東京リージョンで行いますが、途中手順によってオハイオ リージョンを使いますので、文中の記載に従ってください。

1. IAM ロールの設定

ハンズオンで作成する Lambda 関数が他のサービスの実行に必要となる権限を作成しま

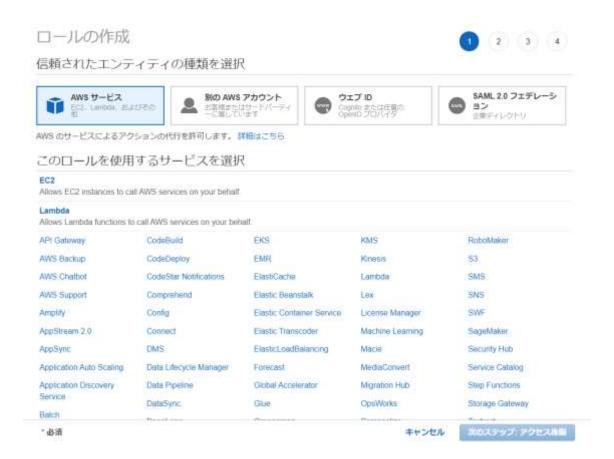
1-1. IAM のトップ画面にいき、「ロール」をクリックします。



1-2. 「ロールの作成」を押します



1-3. 「Lambda」を選択し、「次のステップ:アクセス権限」を押します

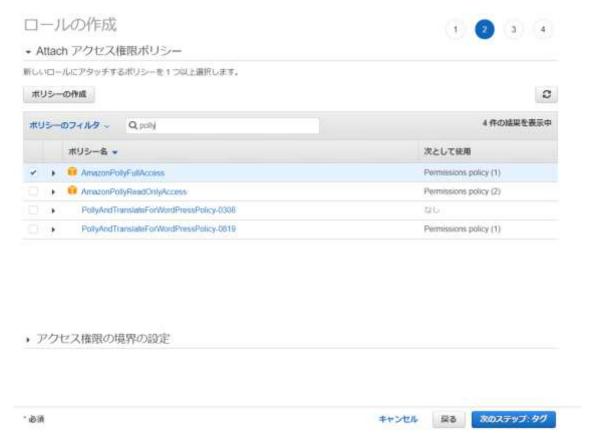


1-4.検索項目に[translate]と入力し、出てきた「TranslateFullAccess」を選びます。右

下のボタンはまだ押しません。



1-4. 同様に「Polly」「Transcribe」「Comprehend」も検索し FullAccess にチェックを入れ最後に、「次のステップ: タグ | を押します。



1-5. 何も入力せず「次のステップ:確認」を押します。



1-6. ロール名に「YYYYMMDDnlphandson」と入力します。(YYYYMMDD は本日の日

付)



1-7. 4つのサービスへの FullAccess がついていることを確認し、「ロールの作成」を押します。以下のような緑で成功が表示されたら作業完了です。皆さんが作成したロールの名前が表示されているはずです。

✓ ロール 20200117 が作成されました。

2. Amazon Translate

#この手順はオプションです。作業を飛ばしても問題ありません

2-1. 何か適当な英語のニュースを探してコピペしておきます。あまり長い文字だと作業

が不必要に大変になるので、だいたい150文字以内にしましょう。

2-2.AWS Lambda の管理者画面にいき、画面右上の「関数の作成」を押します。



2-3.関数名に「YYYYMMDDtranslate」(YYYYMMDDは本日の日付)を入力し、ランタイムは「Python3.8」を選択します。

STAC. POSTABLE SCHERALDER CHAP	1.

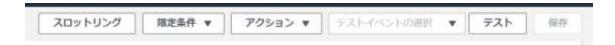
2-4. 「実行ロールの選択または作成」を伸ばして、「既存のロールを選択する」を選び先ほど作成した IAM ロールを選択します。選択がおわったら「関数の作成」を押します。

実行ロール 開始ロードス利用を定義するコームを規則します。カスタムワールを作成するエユニMM コンゲールに作動します。 ■ 基本的な Landob アクセス機能で新しいロールを作成 ① 微砂のロールを使用する ② MS ポリシーテンプシートから取し、ソロールを作成				
を存むロール この Landes 製造で見することに対応して単原のロールを選手します。このロールには、 Procent Case Month Logo 2137をディブロードするアジセスの名字を集です。				
20200117	*	O		
AM ID-O				
			キャンセル	Millionia

2-5.サンプルで挿入されているコードをすべて消し「translate lambda.txt」の内容をコピペします。その後右上の「保存」を押します。



2-6. 右上の「テスト」を押します。



2-7. イベント名に適当な名前を入れあらかじめ挿入されている文字列を削除し、

「translatelambdatest.txt」の中身に置き換えます。



2-8. [hello]の部分をコピペしておいた英語の記事に入れ替えます。この際改行は全て抜いてください。その後右下の「作成」を押します。



2-9. 右のブラウザのスクロールバーを一番上に持っていき、「テスト」を押してください。以下が表示されれば成功です。



2-10. 「詳細」を伸ばすと翻訳された日本語がでてきます。



3. Amazon Polly

先ほどは Lambda で Translate を操作しましたが、実は、Translate、Polly、

Transcribe、Comprehend は全てマネージメントコンソールからテストすることができるようになっています。このハンズオンではこの後何度も Polly を使用しますので、マネージメントコンソールでの手順を紹介します。

3-1. Polly の管理者画面にアクセスします。



3-2. Translate で翻訳した文字列をコピペします。Translate の手順を飛ばした方は、な

んでもいいので150字以下程度の文字列をコピペします。



- 3-3. 男性か女性の音声を選び、「ダウンロード MP3」を押すと、しばらく待てば mp3 ファイルが生成されます。人名や地名など難読文字のカスタマイズもこの画面から可能ですが、このハンズオンでは割愛します。興味がある方は講師に聞いてください。
- 3-4. ダウンロードされた mp3 を再生し、正しく変換されていることを確認します。

4. Transcribe

先ほど生成した mp3 から文字起こしを行う環境を作ります。

4-1. S3 バケットを東京リージョンで作るため、S3 の管理者画面に行きます。



4-2. 「バケットを作成する」を押して YYYYMMDDName (YYYYMMDD は本日の日

付。Name はご自身のアルファベットの名前)を入れます。リージョンが東京となっていることを確認し、「次へ」を押します。



4-3. すべてデフォルトで作成しますので、もう 2 回「次へ」を押し、最後に「バケットを作成」を押します。



4-4. 先ほど作成したバケットをクリックし「アップロード」を押します。



4-5. 先ほど作成した mp3 をアップロードします。



4-6. デフォルトのままアップロードしますので、「次へ」を3回、「アップロード」を1回押してください。完了すると以下のような画面になります。



4-6-1. 先ほど作成した IAM ロールは、Translate、Polly、Transcribe、Comprehend への権限が付与されていますが、このままでは S3 へのアクセスが許可されていません。
IAM ロールの画面に戻り、S3へのアクセス権限を付与します。先ほど作成したロールを
IAM ロール管理画面から特定し、名前をクリックします。



4-6-2. 「ポリシーをアタッチします」をおします

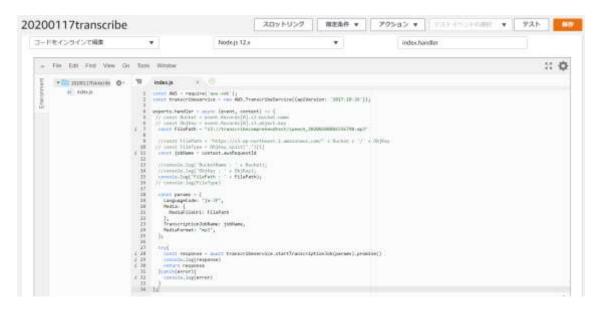


- 4-6-3. 「AmazonS3FullAccess」を選び「ポリシーのアタッチ」を押します。
- 4-7. 別のタブで Lambda の管理者画面を開き、Translate と同じ要領で

YYYYMMDDtranscribe という関数を作ります。今度はランタイムに Node12.x

を選択します。 先ほど作成したロールを指定することを忘れずに行って下さい。

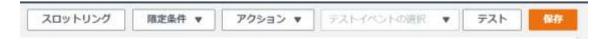
4-8. 「transcribelambda.txt」の中身をコピペします。



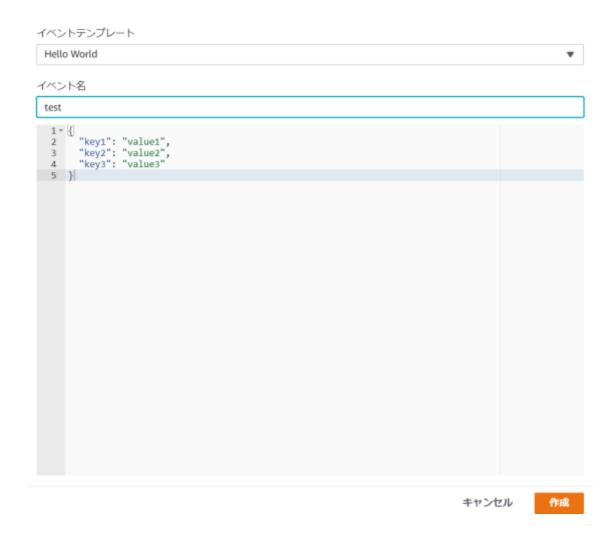
4-9. [const FilePath]の部分を、[先ほど作成したバケット名/アップした mp3 ファイル名]に変換します。以下のような文字列になるはずです。

"s3://20200117kameda/speech_20200117061325085.mp3"

4-10. 右上の「保存」を押します。



4-11. 「テスト」を押します。イベント名に適当な名前を入れて、「作成」を押します。 #この関数では外部インプトは必要ありませんので、文字列はデフォルトのままダミーの テストイベントを作成する必要があります。



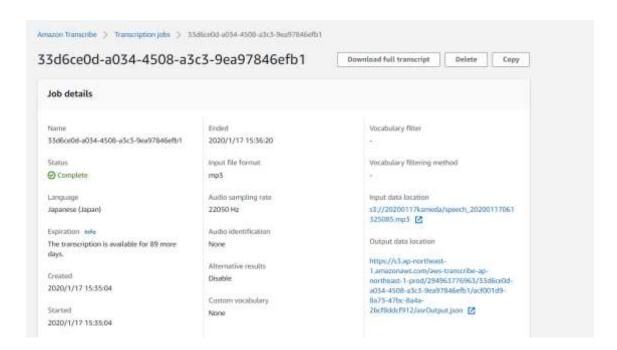
4-12.ブラウザのスクロールバーを一番に持っていき「テスト」を押します。以下のような表示[IN_PROGRESS]が戻れば成功です。

4-13. Transcribe の管理者画面に行きます。プロジェクトが開始されていますので、

[Complete]になるまでしばらくまちます。



4-14. [Complete]になったら、Name をクリックします。



4-15. [Download full transcript]を押すと json がダウンロードされますので、開くと文字お越しされた文字列が格納されています。青い部分を次の Step で使いますので、コピペしておきます。

```
Total series | 1.8 content | 1.1 type | promunciation | 1.2 type | 1.2 type | 1.3 type |
```

5. Comprehend

先ほどの Transcribe が書き起こした mp3 の内容を Comprehend で分析します。

5-1. **2020 年 1** 月現在、Comprehend は東京リージョンにはないため、オレゴンリージョンで関数を作ります。画面上のリージョンをオレゴンに切り替えます。

Lambda 関数の作成画面に再度いき、YYYYMMDDComprehend で関数を作成します。

ランタイムは Python3.8 を指定し、先ほどと同様に今日作成した IAM ロールを設定する ことを忘れずに行います。

5-2. [comprehendlambda.txt]の中身をコピペします。

このサンプルコードは、Comprehend が Negative なメッセージであると判断した場合、

"エスカレーションをしてください"というメッセージが出力されます。



- 5-3. 画面右上の「保存」を押します。その後「テスト」を押し、イベント名に適当な名前を付けた後、[comprehendlambdatest.txt]をコピペします。
- 5-4. テスト文字列には「あなたを攻撃します」と記載されていますが、一度そのままテストを実行します。以下のような画面が出力されます。

- 5-5. 先ほどのテストで使った文字列を「あなたが好きです」に変えて再度テストを実行してください。修正するためには、以下の箇所から「テストイベントの設定」を選択することで、先ほど設定したテストを上書きすることができます。
- 5-6. テストを実行すると分析結果が「Positive」に代わり、エスカレーションがされなくなりました。

```
● 実行結果:成功 (ログ)

▼ 詳細

関数の実行から返された結果が以下のエリアに表示されます。関数から結果を返す方法の詳細については、こちらを参照してください。

{
    "sessionAttributes": {
        "sentiment": "POSITIVE"
},
    "dialogAction": {
        "type": "Delegate",
        "slots": {
        "slots": "None",
        "slotz": "None"
}
```

5-7. では先ほど Transcribe で文字お越しした文字列に入れ替えて再度テストをしてみてください。ほとんどのニュース記事の場合、例え内容が悲惨な事件などであっても、その論調は感情を持たせず客観的に記載されていますので、[Neutral]と出力されるケースが多いようです。

```
● 実行結果:成功 (ログ)

▼ 詳細

開致の実行から返された結果が以下のエリアに表示されます。開致から結果を返す方法の詳細については、ごちらを参照してください。

{
    "sessionAttributes": {
        "sentiment": "NEUTRAL"
    },
    "dialogAction": {
        "type": "Delegate",
        "slots": "None",
        "slot1": "None",
        "slot2": "None"
    }
```

- 6. Transcribe と Comprehend を連携させ、音声の分析結果を Elasticsearch service の kibana で可視化する。以下の作業はすべてオレゴンで行います。Cloud9 は起動した後、 個別の IAM ロール設定が可能ですが、デフォルトでは、AWS の管理者画面にログインしている IAM の権限を引き継ぎます。この手順ではデフォルト権限設定を利用するため、 Comprehend、Transcribe、ElaasticSearch Service、S3、Cloud9 へのフル権限を 持っている IAM ユーザーで管理者画面にログインして作業を行ってください。
- 6-1. Cloud9 を起動するため、Cloud9 の管理者画面にいきます。



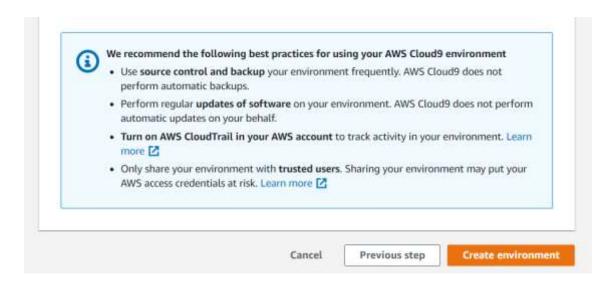
6-2. 「Create Environment」を押します。[Name]には YYYYMMDDhandson (YYYYMMDD は本日の日付) と入力し、[Next Step]を押します。。

	e and description		
Name The name needs to be unio	ue per user. You can update it at any time in you	ur environment settings.	
Name			
	vironment's card in your dashboard. You can up on for your environment	date it at any time in your environment settin	gs.

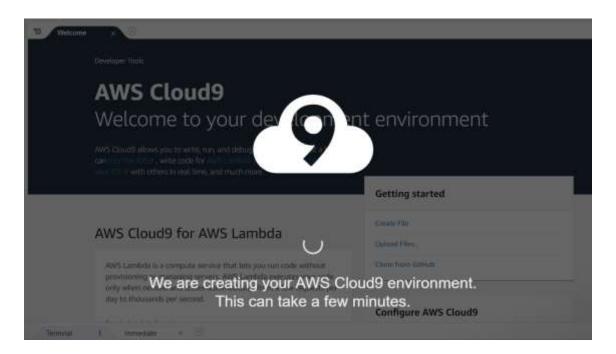
6-3. 画面下の[Network Settings]を以下の様に伸ばし、パブリックサブネットを持つ VPC とパブリックサブネットを選択します。不明な場合は、チューターに聞いて下さい。 別タブで VPC の管理者画面を開くと確認しやすくなります。

	unch your EC2 instance into an existing Amazon Private Cloud	d (VPC) or cr	eate a new o	one.
elect a range of IP addresses in your VPC to isolate EC2 resources from each other.	vpc-0f86d0fce5ac8aca0 (default)		C	☑ Create new VPC
No preference (default subpet in any Availability Zone) ▼ (7 12 Create new subpet	ubnet elect a range of IP addresses in your VPC to isolate EC2 resourc	ces from eac	th other.	
	elect a range of IP addresses in your VPC to isolate EC2 resource			

6-4. 「Next Step」を押します。



6-5. 「Create environment」を押すと以下の画面に切り替わります。



- 6-6. 起動は数分間かかりますが、完了するとコンソールが使えるようになります。
- 6-7. オレゴンに S3 バケットを作成するために、以下のコマンドをタイプして下さい。

aws s3 mb s3://<yyyymmddtranscribename> --region us-west-2

(YYYYMMDD は本日の日付、name はお名前)

- 6-8. 念のため S3 の管理者画面で S3 バケットが正しく作られているか確認します。表示されない場合、再読み込みをしてみましょう。
- 6-9. オレゴンに Elasticsearch service を起動するため Elasticsearch service の管理者画面に遷移します。

Amazon Elasticsearch Service ダッシュボード

新しいドメインの作成

Elasticsearch ドメイン

6-10. 「新しいドメインの作成」を押します。

デプロイタイプの選択

デプロイタイプはユースケースの一般的な設定を指定します。これらの設定は、ドメインの作成後に、いつでも変更できます。

デプロイタイプ		本番稼働用 受走したパフォーマンスを確保する場合、模倣のアベイラビリティーゾーンと専用マ スターノード。
	•	開発およびテスト Elasticsearch エンドポイントが 1つのみ必要な場合。1つのアペイラビリティーゾー ン。
		カスタム ▼ベての利用可能なオプションから設定を選択します。
		UltraWarm プレビュー 選択して UltraWarm ミプレビューします。

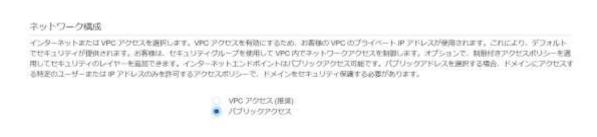
6-11.「開発およびテスト」を選びます。



- 6-12. 「6.2」を選んで「次へ」を押します。
- 6-13. ドメイン名に[transcribetestYYYYMMDD](YYYYMMDD は今日の日付)と入力します。



- 6-14. 「次へ」を押します。
- 6-15. ネトワーク構成で「パブリックアクセス」を選びます。



6-16. 「アクセスポリシー」を以下のようにして Public の Anonymous アクセスが可能な環境を作ります。

アクセスポリシー				
アクセスポリシーは、Amazon Elasticsearch Service のド ー、またはロールを指定する場合。リクエストに署名する			るかを解御します。この	ポソシーでアカウント・ユーザ
カスタムポリシービルダーでは、 雌大 10 個の要素を許可	します。JSON 定義のア	クセスポリシーを使用して、10 個	性部える要素を持つボル	/シーを定義します。
ドメインアクセスポリシー	カスタムアクセスポリ	<i>15</i> − •		
	WWS アカウントID、アカ スを許可または物面します	SOF ARM, NAM IT - IT - ARM, NAM	D-JLARN, IFVE PFL-X	・または CDR プロックによるアクセ
	₽4.アドレス ▼	© .	\$100 -	费惠在研除
	IAM ARN	3	14回 -	數萬を胡除
	更素和iBml			

6-17. 暗号化はそのまま設定します。



- 6-18. 「次へ」を押します。最後に確認画面が表示されるので「確認」を押します。
- 6-19. 現在作成中です。作成が完了したらエンドポイント及び Kibana の URL が表示されますので、しばらくまち、コピーしておきます。(別のタブで Cloud9 の画面に戻って次の Step に進んでも問題ありません)



6-20. 以下 2 つのコマンドを Cloud9 で実行します。コピペすると正常に動作しないケースがありますので。手で入力します。

python --version

pip --version

ハンズオンを行う時期によって出力が異なりますが、以下のように Python はバージョン3、pip がバージョン 2 を指している場合、以下のコマンドを実行して、pip の向き先を変更します。

sudo update-alternatives --config python

Selection	Command
*+ 1 2	/usr/bin/python2.7 /usr/bin/python3.6
Enter to keep	the current selection[+], or type selection number:

2を選びます。これで Python3 環境の整備が終わりました。

6-21. 必要なライブラリをインストールするため、以下の3つのコマンドを実行します。赤字で「Permission denied」と表示された場合、先頭に sudo をつけて実行してください。

pip install boto3

pip install requests

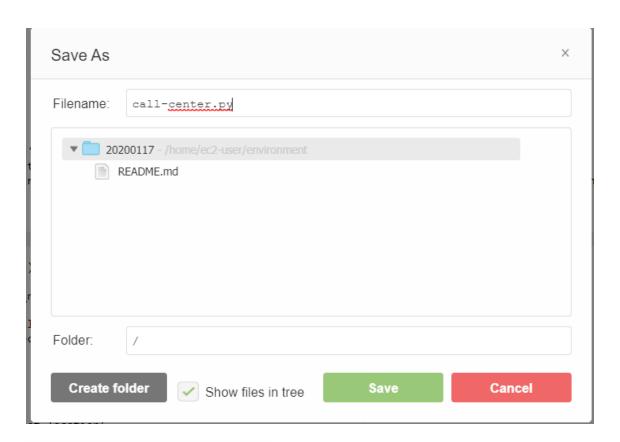
pip install requests_aws4auth

(エラーとなる場合は、sudo pip install requests_aws4auth としてください)

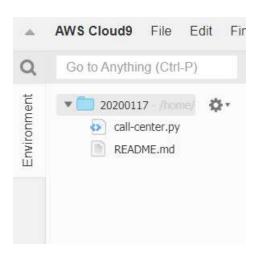
6-22. 必要なライブラリのインストールが完了しましたので、File→New File と選んで、新しく表れたタブに call-center.py の中身をコピペします。



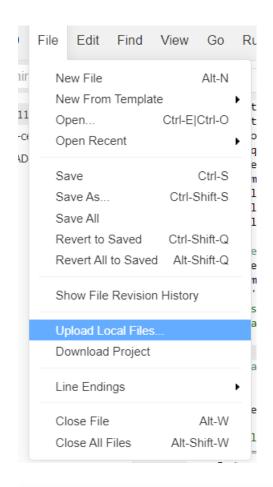
6-23. File→Save 通して、Filename に[call-center.py]と入力し「Save」を押します。



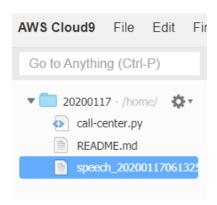
6-24. 以下のフォルダ構成となります。



6-25. File→Upload Local Files を選び、先ほど作成した mp3 をアップロードします。



6-26. Upload 画面は自動で閉じませんので、右上のバツを押して閉じます。以下のフォルダ



構成となります。

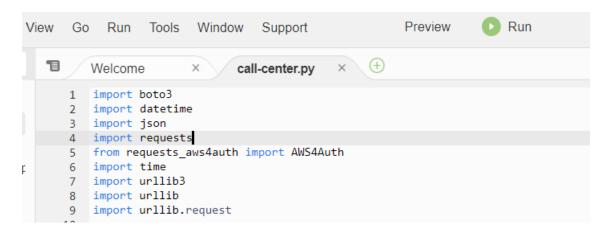
6-27. ソースコードを修正します。12 行目、13 行目、14 行目を修正します。

12 行目: アップロードした mp3 ファイル名

13 行目: オレゴンに作成した S3 バケット名

14 行目: Elasticsearch service のエンドポイント

6-28. ファイルを File→Save したらいよいよ実行です。 画面上の Run ボタンを押します。



6-29. 以下の画面が表示されれば正しく動作しています。



6-30. オレゴンリージョンの Transcribe の画面に行きます。文字起こしが開始されていま

す。



#ここで生成される job 名は mp3 ファイル名です。同じファイル名の解析を再実行すると、

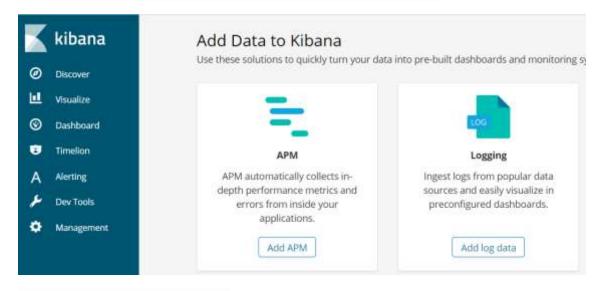
重複エラーが発生する他、再度実行する場合、ファイル名を変更するか、管理者画面から
Complete となった job を削除するかしてください。Python が得意な方は、プログラムを修正し都度乱数や時間をベースとした job 名が生成されるように書き換えることもできます。

6-31. 以下のような表示がされれば実行は完了です Transcribe が文字起こしを行い、その後 Comprehend が Keyword 分析を行い Elasticsearch Service にデータが投入されています。

```
Still matting .

Still
```

6-31. 先ほどコピーした Kibana の URL にブラウザでアクセします。



6-32. 「Visualize」を押します。



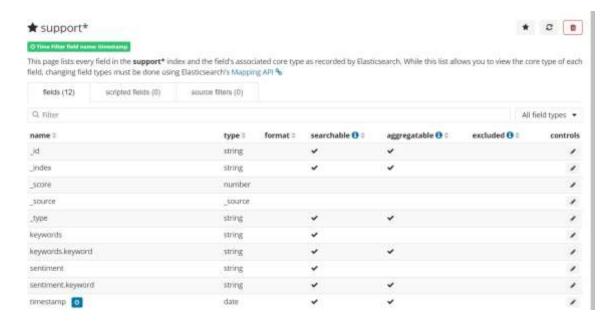
6-32. [Index Pattern]に[support*]と入力し[Next Step]を押します。

> Next sten

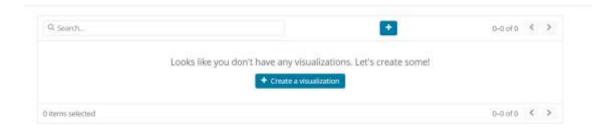
6-33. [Time Filter field name]に timestamp を選び[Create Index Pattern]を押します。

na uses index patterns to retneve d	ata from Elasticsearch indices for things like vi	ASSAURZATIONS.
Step 2 of 2: Configure settings		
You've defined support* as your inc	tex pattern. Now you can specify some setting	gs before we create it.
Time Filter field name	Refresh	
timestamp	· ·	
the Time #8ter will use this field to filter yo	our data by time;	
rou can choose not to have a time field, bu sarrow down your data by a time range.	at you will not be able to	
> Show advanced options		
		⟨ Back

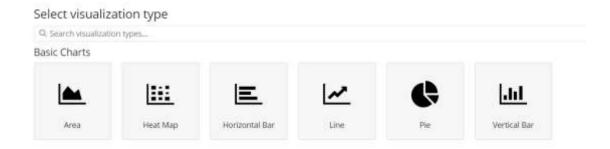
6-34. 正しくデータが投入され、設定も正しければ以下の画面が表示されます。



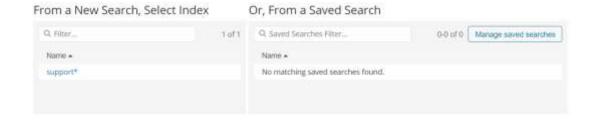
6-35. もう一度 Visualization を押します。



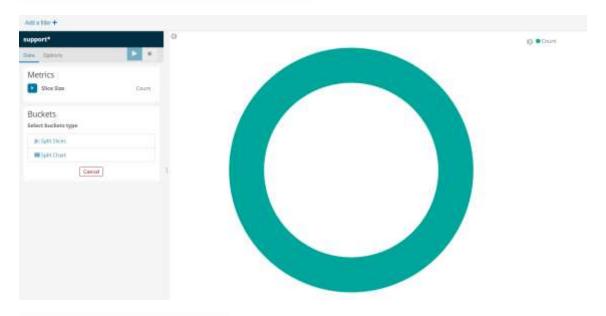
6-36. [Create a visualization]を押します。



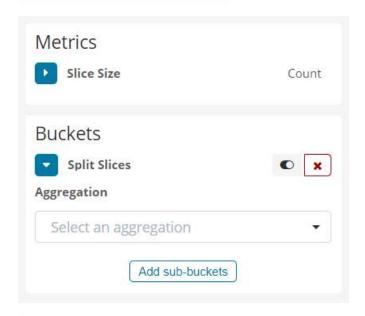
6-37. 円グラフを選びます。



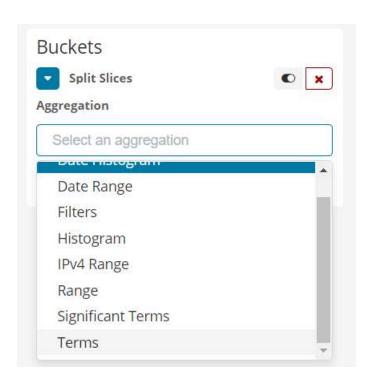
6-38. 先ほど作成した[support*]を選びます。



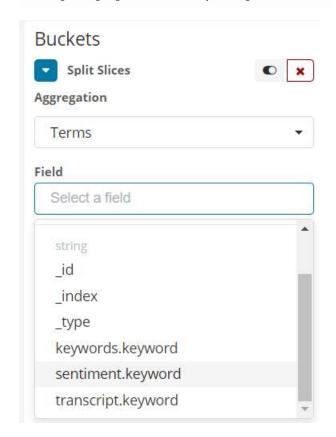
6-39. [Split Slices]を選びます。



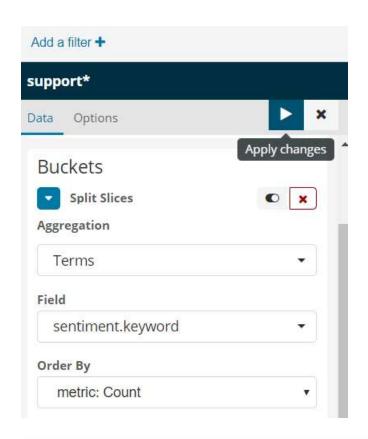
6-40. [Aggregation]で[Term]を選びます。



6-41. [Filed]で[sentiment.keyword]を選びます。



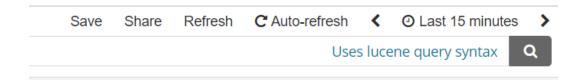
6-42. [apply changes]を押します。



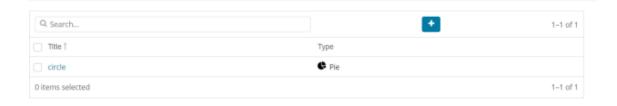
6-43. 先ほど分析した文字列の結果 (Positive, Neutral, Negative) が表示されます。まだ 1 件しか分析結果がありませんが、複数分析を行えば増えていきます。



6-44. 右上の[Save]で適当な名前で保存しておきます。



6-45. 再度[Visualization]を押します。

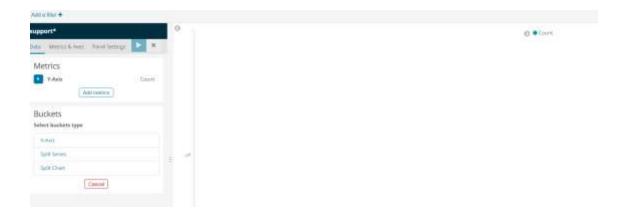


6-46. 先ほどのものが保存されていますが、もう1つ作成するので+ を押します。

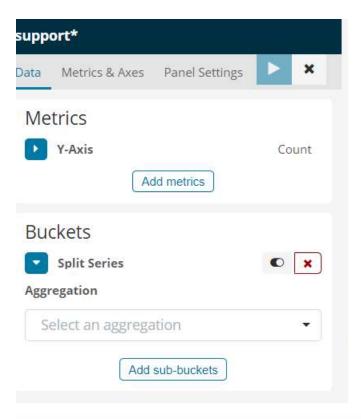
6-47. 今度は Horizontal Bar を選びます。



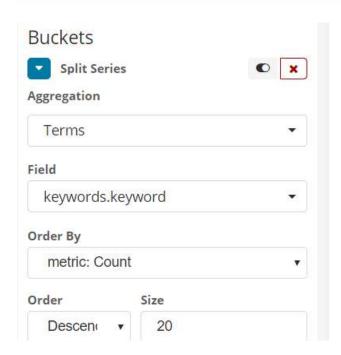
6-48. また[support*]を選びます。



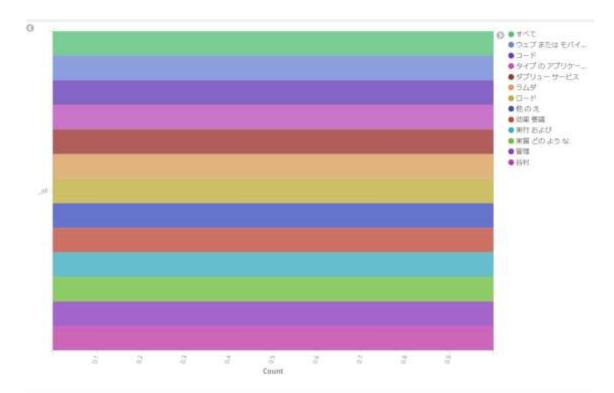
6-49. 今度は[Split Series]を選びます。



6-50. 以下の画面のように設定し、Apply changes を押します。



6-51. 以下のように、取り込んだ単語が件数毎にグラフ化されます。



まだ1件しか取り込んでいないため、全て1件ですが、複数回繰り返すうちにグラフ生成され ていきます。

7. リソースの削除

本日生成した Elasticesearcgh Service, Cloud9, S3 バケットを削除してください。 おつかれさまでした!