テキスト分析
（Text analytics）

テキスト分析（Text analytics）の位置付け

```
Azure AI services
 └ Azure AI Language
   └ Text analytics（テキスト分析）
      ├ Language detection（言語検出）
      ├ Key phraze extraction（キー フレーズ抽出）
      ├ Sentiment analysis（感情分析）
      ├ Named Entity Recognition（NER）（エンティティ抽出）
      ├ Entity linking（エンティティリンキング）
      ├ PII detection（PII 検出）
      └ Summarization（要約）
```

※ PII (Personally Identifiable Information) 検出: 個人を特定できる情報の検出

テキスト分析の概要

- 公式ドキュメント (Azure AI Language)
- テキストを理解し分析するための自然言語処理 (NLP) 機能を提供
- 新機能が追加される場合がある
    - 追加された機能の例
        - ネイティブドキュメントサポート
        - テキスト分析 for health
- 提供されていた機能が「非推奨」（廃止）となる場合がある
    - 廃止済みとなった機能の例
        - カスタム感情分析
            - カスタム テキスト分類（Custom text classification） への移行にて対応。

テキスト分析に含まれる主な機能(1)

- Language detection（言語検出）: テキストの言語を特定
- Key phraze extraction（キー フレーズ抽出）: テキストの主要な概念を抽出
- Sentiment analysis（感情分析）: テキストの感情を分析
  - Opinion mining（オピニオンマイニング）: テキストに含まれる複数の側面についてそれぞれ分析
- Named Entity Recognition (NER)（エンティティ抽出）: テキストから場所、数量、日付などを抽出
  - Entity linking（エンティティリンキング）: 抽出したそれぞれのエンティティにWikipediaへのリンクを設定

テキスト分析に含まれる主な機能(2)

- Personally Identifiable Information (PII) detection（PII 検出）: テキストに含まれる人名・メールアドレス・電話番号などを検出・置換
- Summarization（要約）: 長いテキストを要約して概要を作成
- ネイティブドキュメントサポート: Word、 PDF、テキストファイルなどに対しPII検出や要約を非同期的に実行
- Text analytics for health: テキストから診断、薬剤名、症状などの医療情報を抽出

キーフレーズ抽出

- 入力されたテキスト内の主要な概念を抽出する
- つまり「この文章は何について説明しているか」を抽出
- 例: 「料理はすばらしく、スタッフはみな親切だった」 => 料理、スタッフ

エンティティ抽出 (Named Entity Recognition, NER)

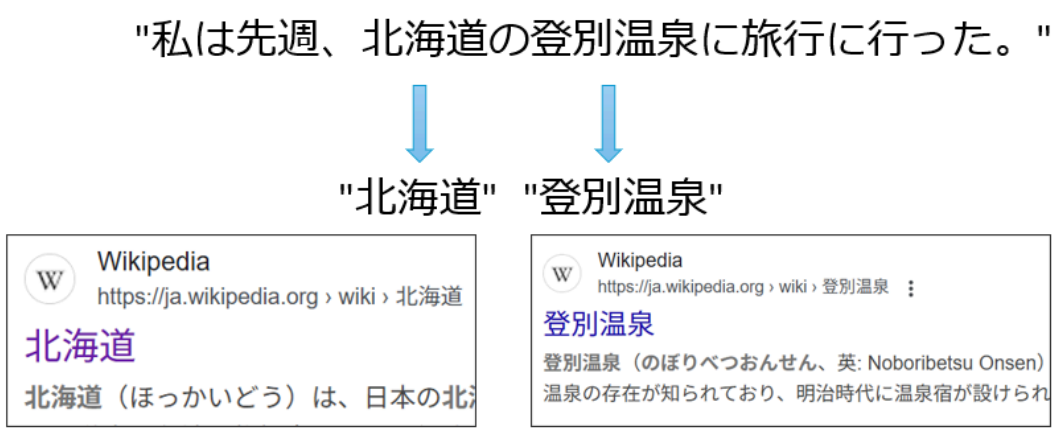
- テキスト内の日時、場所などを特定。

"私は先週、北海道の登別温泉に旅行に行った。"

"先週" (日時)　"北海道" (場所)　"登別温泉" (場所)　"旅行" (イベント)

エンティティリンキング

- テキスト内のエンティティを認識してWikipediaページに関連付けする。



※2025/5現在、エンティティリンキングのリンク先Wikipediaとしては英語版とスペイン語版のみ対応。上記の図では、学習者がイメージしやすいように、Wikipedia日本語版を表示している。

個人を特定できる情報の検出 - PII (Personally Identifiable Information) detection

- テキストに含まれる電話番号やメールアドレスを特定
- 特定した部分をマスク処理

"私のメールアドレスは yamada@contoso.com です"

"私のメールアドレスは ********** です"

感情分析 - sentiment analysis (1)

- 入力された文の感情を判定
  - sentiment (positive / negative / neutral / mixed)、 positive_score、negative_score、 neutral_scoreが返される
- 「ポジティブ」（肯定的）
  - 例: 「山田先生の講義はわかりやすい」 => sentiment = positive, positive_score = 0.97
- 「ネガティブ」（否定的）
  - 例: 「山田先生の講義はわかりにくい」 => sentiment = negative, negative_score 0.85

感情分析 - sentiment analysis (2)

- 「ニュートラル」（中立的）
  - 例: 「吾輩は猫である」 => sentiment: neutral, neutral_score: 0.91
- 「ミックス」（肯定的な意見と否定的な意見が両方含まれている）
  - 例: 「料理はすばらしかった。スタッフは不親切だった」 => sentiment = mixed, positive_score: 0.51, negative_score: 0.49

参考: analysis と analytics の違い ※Azure特有の用語としてではなく一般知識として

```
Azure AI Language
 └ Text analytics（テキスト分析）
    └ Sentiment analysis（感情分析）
```

- 感情 **分析** = sentiment **analysys**
- テキスト **分析** = text **analytics**

用語	意味	用途・ニュアンス
analysis アナリシス	ある対象を詳細に調べて理解する行為。analyze（分析する）という動詞の名詞形	一回限り、または特定の目的のための「分析」行為。
analytics アナリティクス	データを継続的に収集・処理・分析する技術やプロセス	「分析を自動化・体系化」すること。例：顧客行動の分析。

オピニオンマイニング

- 「感情分析」よりも高度な分析機能
- 「料理はすばらしかったがスタッフは不親切だった」といった文章を複数の「主題」に分けて分析する
-

主題(subject)	**意見(opinion)**	**感情(sentiment)**
料理	すばらしい	肯定的(positive)
スタッフ	不親切	否定的(negative)

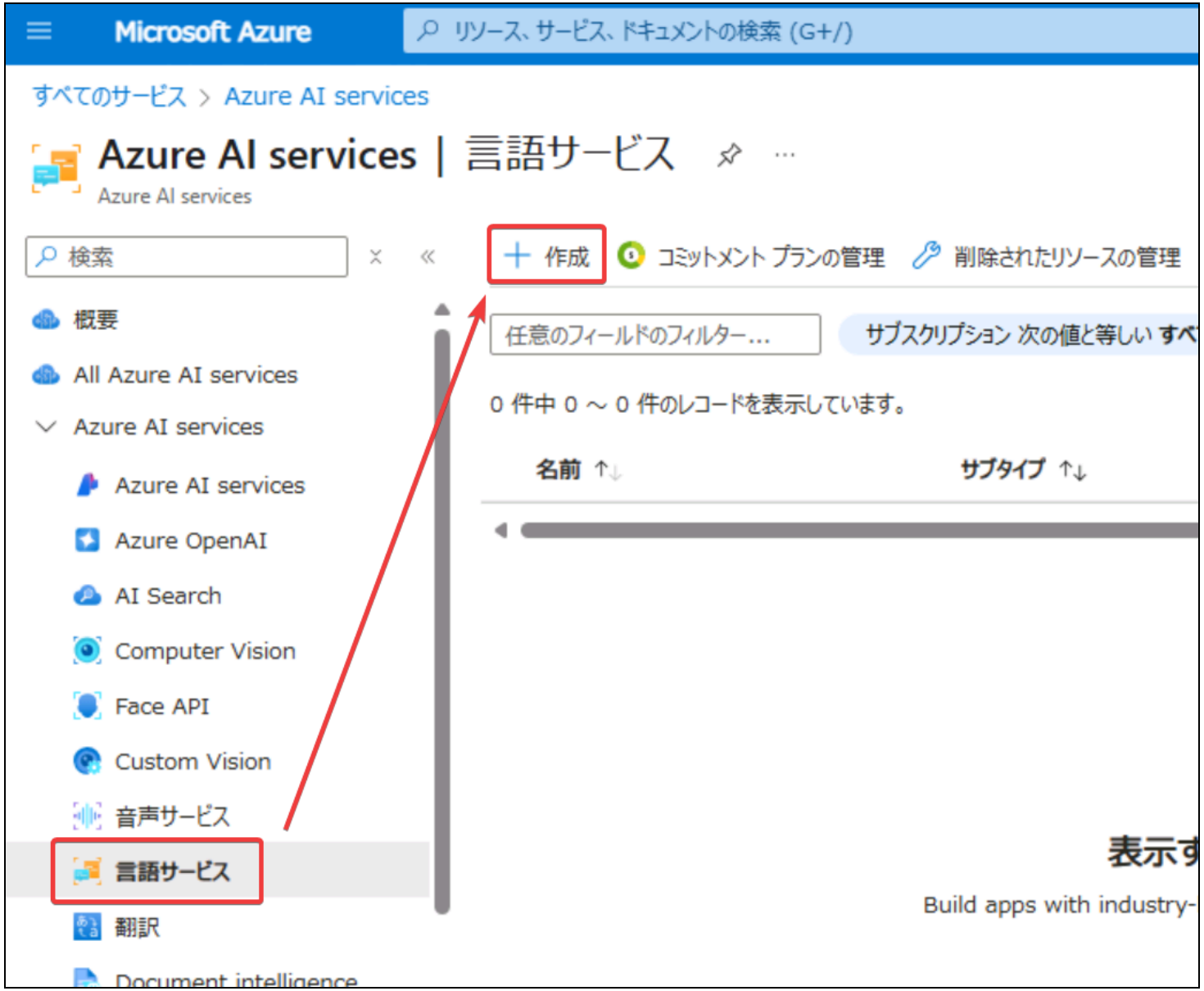
- 各「主題」ごとに「意見」と「感情」が得られる

テキスト分析を行うためのAzureリソース

テキスト分析には以下のいずれかのリソースが必要。

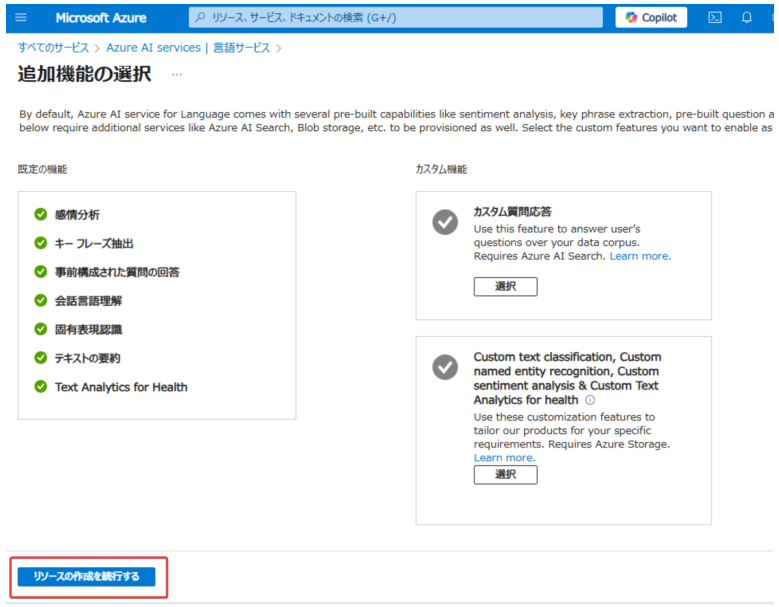
- 「Azure AI Language」リソース
  - 価格レベル: Free (F0) / Standard (S1)
- 「Azure AI services マルチサービスアカウント」リソース
  - 価格レベル: Standard (S1)
- 「Azure AI services」リソース
  - 価格レベル: Standard (S1)

リソースの作成時の注意



リソースの作成時の注意

「追加機能の選択」画面が出てくるが、そのまま「リソースの作成を続行する」をクリック。



※たとえば、抽出できるエンティティの種類をカスタマイズする「カスタムエンティティ抽出」を使用したい場合、画面右側「選択」ボタンをクリックしてからリソースを作成する。

テキスト分析の料金 (1)

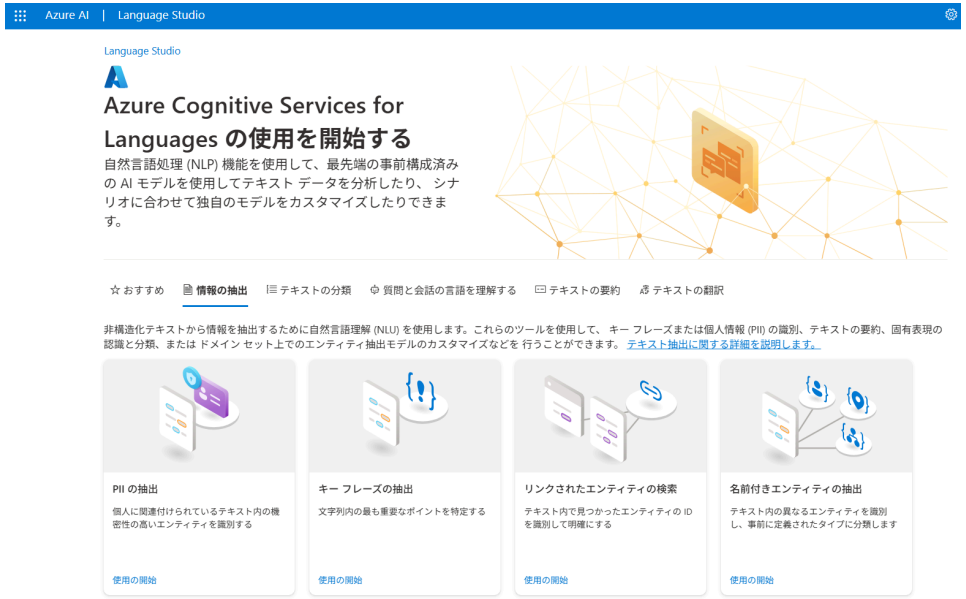
- 価格のページ(Azure AI Language)
- ドキュメント
  - 1回のAPI呼び出しで複数の「ドキュメント」を処理できる。たとえば「はい」と「yes」を「言語検出API」に送信すると2つの「ドキュメント」としてカウント
- テキストレコード
  - ドキュメントそれぞれについて、1,000文字を「テキストレコード」としてカウント
    - 500文字の「ドキュメント」と1,200文字の「ドキュメント」を送信した場合、ドキュメントごとにレコードをカウントし、合計3レコード

テキスト分析の料金 (2)

- 価格レベル Free (F0) のリソース
  - 無料で利用できるが、1 か月あたり 5,000 テキストレコードまで（クォータ）といった制限がある
- 価格レベル Standard (S1) のリソース
  - クォータ制限なし。より高いレートでの利用が可能
  - 機能により単価が異なる
    - 例:「感情分析」: $1/1,000 テキストレコード など

Language Studio

- Azure AI Languageの機能を簡単に試すことができるサイト。
- https://language.cognitive.azure.com/



テキスト分析のための公式ライブラリ - Text Analysis SDK

- SDK = Software Development Kit
- C# / Python / Java / JavaScript の4言語のSDKが提供されている。
- 各機能の「クイックスタート」で、リソースの作成、プロジェクトの作成、SDKの導入、コードの記述と実行などの具体的な手順を確認できる
  - 言語検出
  - キーフレーズ抽出
  - （など、各機能に対するクイックスタートが提供されている。詳細は公式ドキュメントのサイトを参照）

SDKの利用例(C#)

```
// エンドポイントとキーを環境変数から読み取り
string endpoint = Environment.GetEnvironmentVariable("LANGUAGE_ENDPOINT");
string key = Environment.GetEnvironmentVariable("LANGUAGE_KEY");

// クライアントを作成
var client = new TextAnalyticsClient(new Uri(endpoint), new AzureKeyCredential(key));

// 言語の検出を実行
DetectedLanguage detectedLanguage = client.DetectLanguage("Ce document est rédigé en Français.");

// 検出された言語の情報を表示
Console.WriteLine($"Language Name: {detectedLanguage.Name}");
Console.WriteLine($"ISO-6391 Name: {detectedLanguage.Iso6391Name}");
```

実行結果例

```
Language Name: French
ISO-6391 Name: fr
```