



chiTra ハンズオン on AWS

2022/12/03



# chiTra ハンズオン

はじめに

# ハンズオンの進め方

- 本資料は compass の本ハンズオンページにて配布しています
  - <https://worksapplications.connpass.com/event/259016/>
  - コマンドやアクセスキーなどについては、適宜コピー＆ペーストしてください
- 質問などについては、口頭・Zoomのチャットで受け付けます。
  - 遮っていただいて大丈夫です！
  - 個別にルームを作ったの対応も可能です

# SageMaker Studio Lab

## アカウントの作成

- <https://studiolab.sagemaker.aws/> にアクセス
- Request account からアカウントを申請
- referral code 欄に以下のコードを入力
  - 「\*\*\*\*」

The image shows a browser window displaying the SageMaker Studio Lab website. The main heading is 'Learn and experiment machine learning'. Below it, there's a description: 'Quickly create data analytics, scientific machine learning projects with notebook browser.' There are two buttons: 'Request free account' and 'Watch video'. A red arrow points to the 'Request free account' button with the text '1. アカウント申請 ↑'. To the right, the 'Request account' form is shown. It includes fields for 'Enter your email\*', 'Enter your first name', 'Enter your last name', 'Select your country', 'Enter your company or organization name', 'Select your occupation', 'Why are you interested in Amazon SageMaker Studio Lab?', and 'Enter referral code'. A red arrow points to the 'Enter referral code' field with the text '2. リファラルコードを入力 →'. At the bottom of the form is a 'Submit request' button.



# Sudachi・chiTraについて

Sudachi, chiTra とは何か

# Sudachiとは

- 日本語形態素解析器
  - 自然文を形態素(~= 単語)に分割する
  - テキストを計算機で扱う第一歩
- 特徴
  - 大規模・定期的に更新される辞書
  - 複数の分割単位を併用可能
  - 機能のプラグイン化
  - 同義語辞書との連携

# 自然文を計算機で扱う

- テキストを計算機でそのまま扱うのは難しい
  - 何らかの手法で数値列(ベクトル)に変換してから利用する
  - 例えば:
    - 形態素解析の結果を元に各単語の出現回数をカウントする

「すももももももものうち」



...	
1	うち
...	
1	すもも
...	
2	もも
...	

# 言語処理に適したベクトル表現とは

- どんなベクトルに変換するとよいか？
  - 変換元の語や文の関係や意味を推定できるようにしたい
  - サイズが大きくなりすぎないようにしたい
  - 同じ語でも文脈を考慮して調整したい

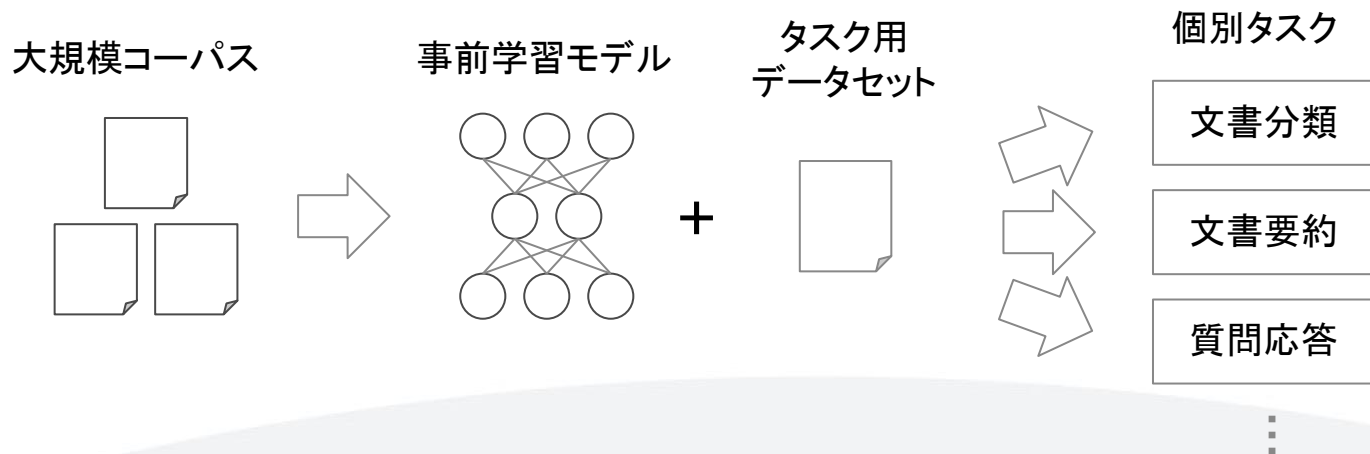
「Appleから発売された新型の マック PCを買った。」 → 「パソコン」

「マックのポテトは揚げたてがおいしい。」 → 「モスバーガー」  
「ロッテリア」



# 事前学習モデル

- 近年はニューラルモデルで変換を行うのが一般的
  - 大量の文書(コーパス)から意味を事前に学習しておく
  - 応用したいタスクに合わせてモデルを変形し、end-to-endで学習させられる



chiTra: Suda~~chi~~ Transformers [tʃi: tara]

- 事前学習済みの大規模ニューラル言語モデルおよびトークナイザ
- Sudachi と Hugging Face Transformers (ニューラル言語モデルのフレームワークを連携させる)

ニューラル言語モデルでは、文章をトークン (モデル用の単語) に分割して扱うのが一般的

- → Sudachi による分割の単位で入出力を扱えるように

chiTra モデル (v1.0) は事前学習済みの BERT モデル

- ファインチューニングすることで様々なタスクに応用可能
- 国立国語研究所の大規模コーパス NWJC で学習
- Sudachi の正規化情報を利用し表記ゆれに頑健
- Apache 2.0 ライセンスで一般公開、商用利用も可能



# chiTra ハンズオン

作業の流れ

# ハンズオンの流れ

本ハンズオンは大きく以下のステップで進めていきます。

- 作業環境を準備する
  - SageMaker Studio Lab を作業環境として整えます
  - AWSにアクセスするためのユーザを設定し、ハンズオンで使用するリソースをダウンロードします
- Sudachi と chiTra を使う
  - Sudachi と chiTra をPython言語で使します
  - 上でダウンロードした Jupyter ノートブックで作業を行います
- chiTra モデルをファインチューニングする
  - モデルをタスクに合わせて調整するファインチューニングを実行します
    - livedoor ニュースの記事分類を題材とします
  - ノートブックから SageMaker のモデル訓練・デプロイ APIを使します

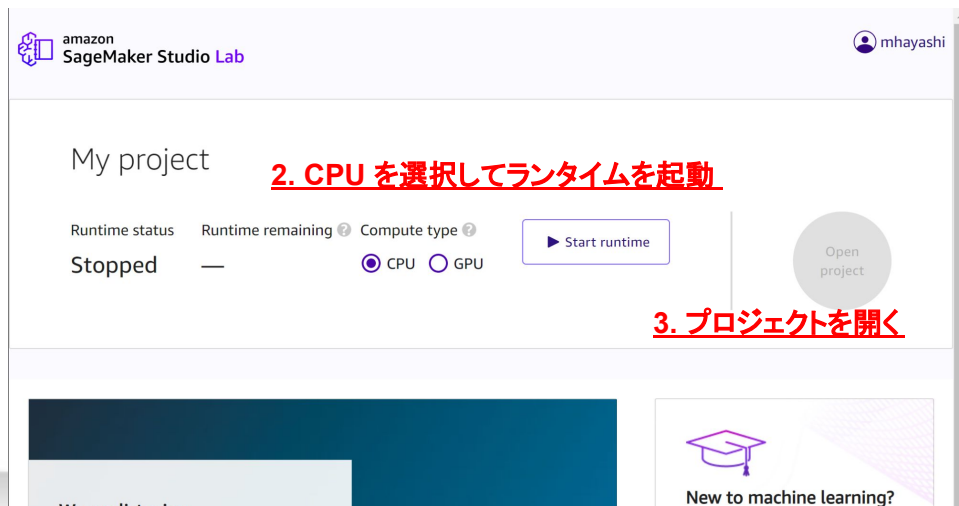
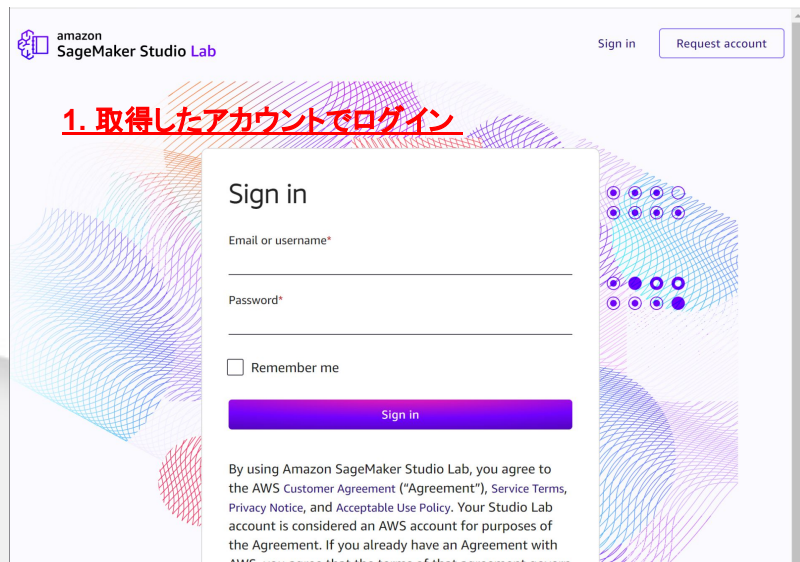
# 作業環境を準備する

SageMaker Studio Lab

# SageMaker Studio Lab

## ログインとランタイムの起動

Amazon SageMaker Studio Lab のプロジェクト画面を開きます。  
ランタイムの起動時に電話番号での認証が必要です。

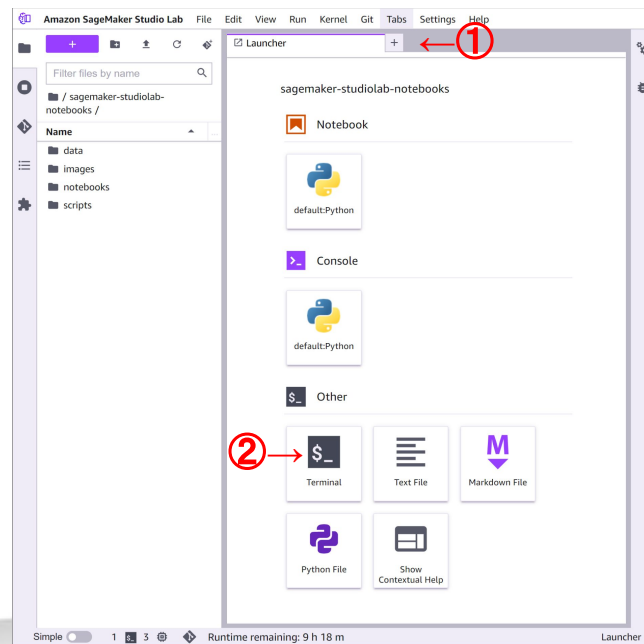


## ターミナルの起動

プロジェクト開始時は右のような画面になります。  
大枠で左側にファイル一覧、右側に開いたファイルが表示されます。

準備ではターミナルを使います。

- Launcherが開いていなければ、上部の＋マークからタブを追加
- 下部にある“Other” -> “Terminal” をダブルクリックして開く



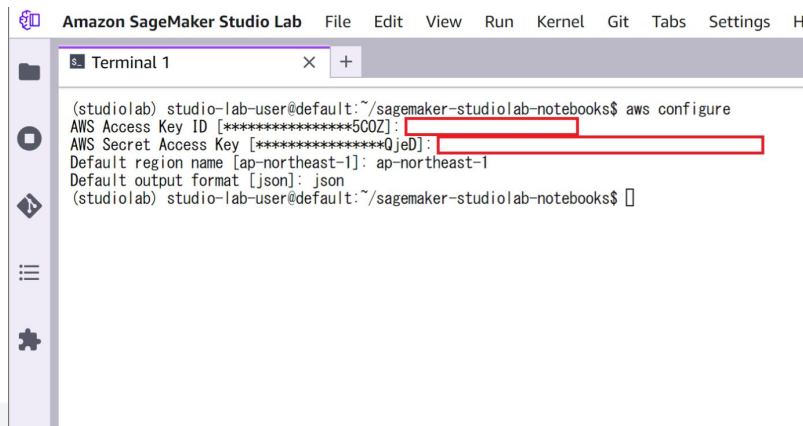
## ユーザの設定

ノートブックからAWSの機能を利用できるようにするため、AWSユーザを設定します。

先ほど開いたターミナルにて`aws configure`と入力し、設定を始めます。下記の内容を入力してください

- Access Key
  - `\*\*\*\*`
- Secret Access Key
  - `\*\*\*\*`
- region: `ap-northeast-1`
- output format: `json`

※アクセスキーは本ハンズオンでのみ使用し、また外部に伝えないようご配慮をお願いします。



```
Amazon SageMaker Studio Lab  File  Edit  View  Run  Kernel  Git  Tabs  Settings  H
Terminal 1
(studiolab) studio-lab-user@default:~/sagemaker-studiolab-notebooks$ aws configure
AWS Access Key ID [*****5C0Z]: 
AWS Secret Access Key [*****QjeD]: 
Default region name [ap-northeast-1]: ap-northeast-1
Default output format [json]: json
(studiolab) studio-lab-user@default:~/sagemaker-studiolab-notebooks$
```



## リソースのダウンロード

本ハンズオンで使用するノートブックや chiTra のモデルデータ等をダウンロードします。

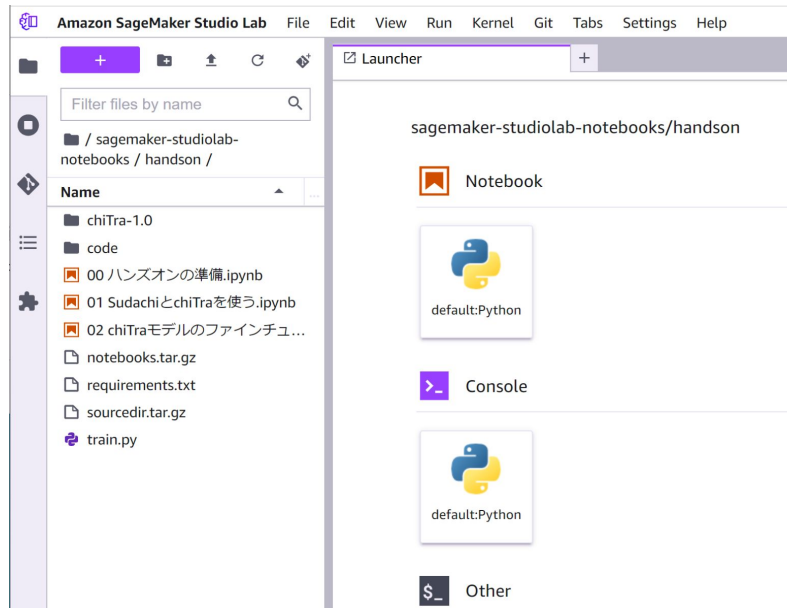
引き続きターミナルから以下を順に入力します。

- s3 から “handson” フォルダへダウンロード
  - ``aws s3 cp --recursive s3://chitra-handson-20221203/source handson``
- “handson” フォルダに移動して解凍
  - ``cd handson``
  - ``tar -xvf sourcedir.tar.gz``
  - ``tar -xvf notebooks.tar.gz``

## 配布物の確認

画面左から "handson" フォルダを開きます。

- 以下のファイルがあることを確認します
  - `00 ハンズオンの準備.ipynb`
  - `01 SudachiとchiTraを使う.ipynb`
  - `02 chiTraモデルのファインチューニング.ipynb`
  - `chiTra-1.0/` フォルダ
  - `code/` フォルダ
  - `train.py`
  - `requirements.txt`



## ライブラリのインストール

ノートブック `00 ハンズオンの準備.ipynb` をダブルクリックして開きます。

実際の画面を共有しながら説明します。


# SudachiとchiTraを使う

Sudachi, chiTra をPythonで呼び出す

# SudachiとchiTraを使う

ノートブック `01 SudachiとchiTraを使う.ipynb` をダブルクリックして開きます。

実際の画面を共有しながら説明します。



# chiTraモデルの ファインチューニング

実タスクへの応用

# chiTraモデルのファインチューニング

ノートブック `02 chiTraモデルのファインチューニング .ipynb` をダブルクリックして開きます。

実際の画面を共有しながら説明します。

# chiTraモデルのファインチューニング

## 訓練ジョブに渡すリソース

訓練ジョブを実行するにあたって、以下のファイルをリソースとしてインスタンスに渡しています。  
これらについて説明します。

- ``chiTra-1.0/``
- ``code/``
- ``requirements.txt``
- ``train.py``



# chiTraモデルのファインチューニング

## 訓練ジョブに渡すリソース

- `chiTra-1.0/`
  - chiTra のデータ
  - 公式 github からダウンロードできるものと同じ
    - <https://github.com/WorksApplications/SudachiTra>

# chiTraモデルのファインチューニング

## 訓練ジョブに渡すリソース

- ``code/``
  - モデルをデプロイする際に使用されるコード
    - 訓練後のモデルと一緒にまとめておく
  - ``requirements.txt``
    - 必要なライブラリを指定
  - ``inference.py``
    - モデルの読み込み・入力テキストの加工の方法を指定
- Hugging FaceのSageMaker連携ドキュメントもご参照ください
  - <https://huggingface.co/docs/sagemaker>

# chiTraモデルのファインチューニング

## 訓練ジョブに渡すリソース

- ``requirements.txt``
  - 訓練に必要な追加ライブラリを指定する
- ``train.py``
  - 訓練インスタンス内で実行されるスクリプト
    - 訓練の本体
  - コードをノートブックに移せば Studio Lab 内でも訓練可能
    - パラメータの受け渡しなどは書き換えが必要
    - 計算量が大きいため、GPUランタイムの使用を推奨



# 終わりに

本日の振り返り

本ハンズオンでは以下を行いました。

- Sudachi, chiTraを使う
  - テキストを形態素に分割する
  - テキストをモデル入力用のトークン列に変換する
  - マスクされた語をモデルで予測する
- chiTraモデルのファインチューニング
  - タスクデータを確認・加工する
  - SageMaker の訓練ジョブを作成・実行する
  - ファインチューニングによるモデル性能の向上を確認する
- モデルのデプロイ
  - 訓練したモデルを SageMaker でデプロイする
  - テキストを渡してラベルを判定させる

# 配布したAWSアカウントについて

本ハンズオンにて配布した AWSアカウントは、近日中に削除されます。

データを以後もご利用になりたい場合はお早目にダウンロードをお願いします。

準備の際に利用したターミナルから以下を実行することで、Studio Labのストレージ内に保存されます。

- 訓練済みモデルのデータ
  - ``aws s3 cp --recursive s3://chitra-handson-20221203/trained/訓練ジョブの名称 model/``
- 訓練に使用した livedoorニュースコーパスの加工済データ
  - ``aws s3 cp --recursive s3://chitra-handson-20221203/datasets/livedoor livedoor/``

# ご参加ありがとうございました

ワークスアプリケーションズでは過去に開催したイベントの資料を SpeakerDeckで公開しています。  
Sudachi や chiTra の機能に関するものもございますのでぜひご覧ください。

- Connpass サイトの弊社グループ → 資料 にリンクがあります
  - <https://speakerdeck.com/waptech>

# Sudachi スポンサーのお願い

ワークスアプリケーションズでは Sudachiの研究開発力を強化し、OSSとしての持続的開発を実現するため、Sudachi開発スポンサーを募集しています。

GitHubスポンサー制度を利用して行っており、各種のリワードをご用意しております。  
詳細は下記リンクよりご覧ください。

<http://nlp.worksap.co.jp/>





## 免責

### ■免責事項および権利帰属について

- ・本資料に関する一切の権利は弊社に帰属します。
- ・本資料には弊社の機密情報が含まれていることがあります。したがって、書面による事前の承諾なしにこれを転載し、または第三者に開示することを禁止いたします。
- ・本資料はディスカッション目的で作成されたものであり、貴社との協議に基づき適宜変更することを想定しております。したがって、弊社は本資料に記載の内容について法的責任を一切負担いたしません。なお、弊社および貴社の法的関係は、今後弊社および貴社が捺印の上締結する契約書に依拠します。弊社は貴社との間で締結された契約書に明示的に記載された責任以外の責任は負担いたしません。
- ・ワークスアプリケーションズ、「HUE®」および「ArielAirOne®」は(株)ワークスアプリケーションズの日本国内における商標または登録商標です。
- ・ワークスアプリケーションズ・エンタープライズは(株)ワークスアプリケーションズ・エンタープライズの日本国内における商標です。
- ・ワークスアプリケーションズ・フロンティアは(株)ワークスアプリケーションズ・フロンティアの日本国内における商標です。
- ・ワークスアプリケーションズ・システムズは(株)ワークスアプリケーションズ・システムズの日本国内における商標です。
- ・本資料に記載された各会社名あるいは各製品名は各社の登録商標または商標です。
- ・本文中および図中では®マークは表記しておりません。
- ・「ArielAirOne®」は「ArielAirOne® Enterprise」「ArielAirOne® Portal」および「ArielAirOne® Framework」を含む総称です。