

人文社会ビジネス科学学術院 ビジネス科学研究群 2023年度 春C

テキストマイニング

day 2

スケジュール

day 1

- 講義 — テキストマイニング概説 (津田先生)
- 講義 — 自然言語処理の最新動向

day 2

- 講義 — テキストマイニングの手順
- 演習 — テキスト解析 (1)
- 演習 — データ理解

day 3

- 演習 — テキスト解析 (2)
- 講義&演習 — データ分析 (使い方編)

day 4

- TextMining Studio の紹介
- 講義&講義 — データ分析 (実践編)

day 5

- 講義&講義 — データ分析 (実践編)

(前回) day 1 – レポート課題

- 以下を PDF ファイルで提出してください
 - ChatGPT と、Google Bard または Microsoft Bing AI Chat のどちらかまたは両方のツールを使って、以下の質問を試してください
 1. 日本の総理大臣は誰ですか？
 2. アメリカの大統領は誰ですか？
 - 応答内容の違いを観察し、違いの理由を考察(感想も可)を文章で記述してください
- ※ 何らかの事情で2つ以上のツールが試せない場合、本日の講義の感想を文章で記述してください

| レポート形式 | 提出先 | 期限 |
|--------|--------|----------|
| PDF | manaba | 次回～18:20 |

- 【参考情報】
- ChatGPT の始め方: <https://book.st-hakky.com/docs/chatgpt-register/>
 - Google Bard の始め方: <https://www.gizmodo.jp/2023/05/what-is-google-bard.html>
 - Bing AI Chat の始め方: <https://www.sedesign.co.jp/dxinsight/bing-ai-chat#article-2>

ChatGPT vs. Bing AI Chat vs. Google Bard

- 結果の違いは、主に学習データの収集時期と最新情報へのアクセスの仕組みによる

| (2023/6末時点) | ChatGPT | Bing AI Chat | Google Bard |
|-------------|---|---|--|
| 学習データの収集時期 | 2021年9月 | 同左 | 2023年1月? |
| 最新情報へのアクセス | × 2021年10月以降の 情報にアクセスできない※ | ○ Bing 検索の結果をもと に ChatGPT が回答を 生成 | ○ Google 検索を通して アクセス可能 |
| 仕組み | <p>Diagram illustrating the architecture of ChatGPT:</p> <p>1. A question is input into the LLM (Large Language Model).</p> <p>2. The LLM generates a response.</p> | <p>Diagram illustrating the Retrieval-Augmented Generation (RAG) architecture of Bing AI Chat:</p> <p>1. A question is input into the application.</p> <p>2. The application performs query generation (クエリ生成②) to interact with the LLM.</p> <p>3. The application performs network search (ネット検索③) to interact with the Web.</p> <p>4. The application generates the final response (回答文生成④) based on the LLM and search results.</p> | <p>Diagram illustrating the Retrieval-Augmented Generation (RAG) architecture of Google Bard:</p> <p>1. A question is input into the application.</p> <p>2. The application performs query generation (クエリ生成②) to interact with the LLM.</p> <p>3. The application performs network search (ネット検索③) to interact with the Web.</p> <p>4. The application generates the final response (回答文生成④) based on the LLM and search results.</p> <p>Retrieval-Augmented Generation (RAG)</p> |

※ 2023/7/2 OpenAI は 5/12からβ版として提供していたWebブラウジング機能「Browse with Bing」を一時的に停止

演習環境の準備

(再掲) 無償で利用できる機械学習環境

- 近年、機械学習の教育・研究を目的とした研究用ツールが相次いで登場

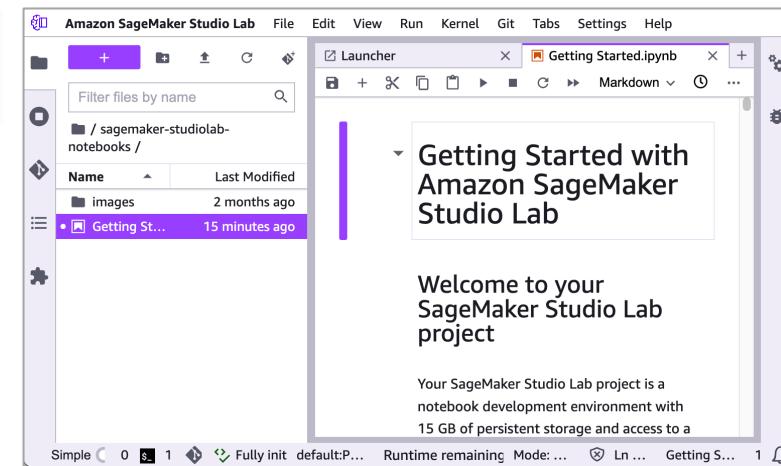
 Colaboratory

<https://colab.research.google.com>



 Amazon SageMaker Studio Lab

<https://studiolab.sagemaker.aws/>



演習で使用
↓

| | Colab(無償版) | Studio Lab |
|-------------|-----------------------|---------------------|
| GPU | T4(16GB) | T4(16GB) |
| 最長実行時間 | 12時間 | CPU:12時間 GPU:4時間 |
| メモリ | 12GB | 15GB |
| ディスク | CPU:100GB GPU:78GB | 15GB (永続化) |
| ターミナル | × | ○ |
| ランタイムの保存と再開 | × | ○ |
| 費用 | 無償 | 無償 |
| その他 | Googleアカウントが必要 | AWSアカウントは不要 (クレカ不要) |

(再掲) SageMaker Studio Lab のアカウント作成

- <https://studiolab.sagemaker.aws/> にログインして、アカウントを作成してください

The screenshot shows the 'Request account' page of the Amazon SageMaker Studio Lab website. The page has a light blue background with abstract line patterns. At the top, there's a logo for 'amazon SageMaker Studio Lab' and a 'Sign in' link. Below that is a large button labeled 'Request account'. The main form contains 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', and 'Why are you interested in Amazon SageMaker Studio Lab?'. At the bottom of the form, there's a yellow-bordered input field labeled 'Enter referral code' and a purple 'Submit request' button.

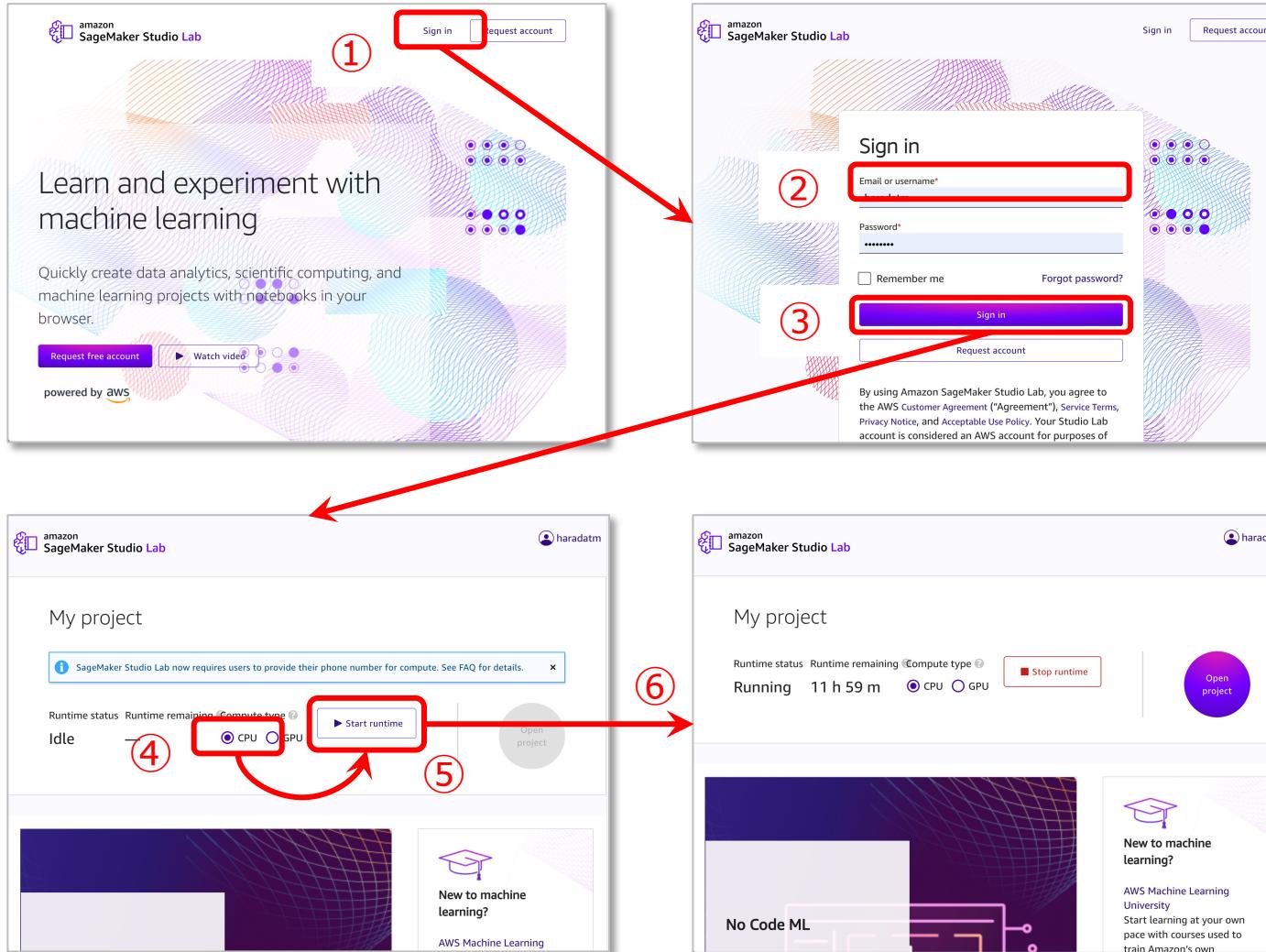
アカウント作成手順

1. [アカウント作成フォーム](#)からアカウントの申し込みを行う
注意: リファラルコードをアカウント作成フォームに忘れずに入力ください
2. 「Account request confirmed ...」のメールを受信し、メール内のリンクからアカウントを作成する
→ リクエストの受付はすぐにメールが届きます
3. 「Verify your email ...」のメールを受信し、メール内のリンクからメールアドレスを認証する
→ リファラルコードを利用している場合は2~3分以内に結果が届きます
4. 「Your account is ready ...」のメールを受信する
→ これで「Sign in」できます

※ リファラルコードの有効期間: 2023/6/30 ~ 2023/7/7

演習環境の準備

- <https://studiolab.sagemaker.aws/> にログインして、Runtimeを起動してください



1. ログインする

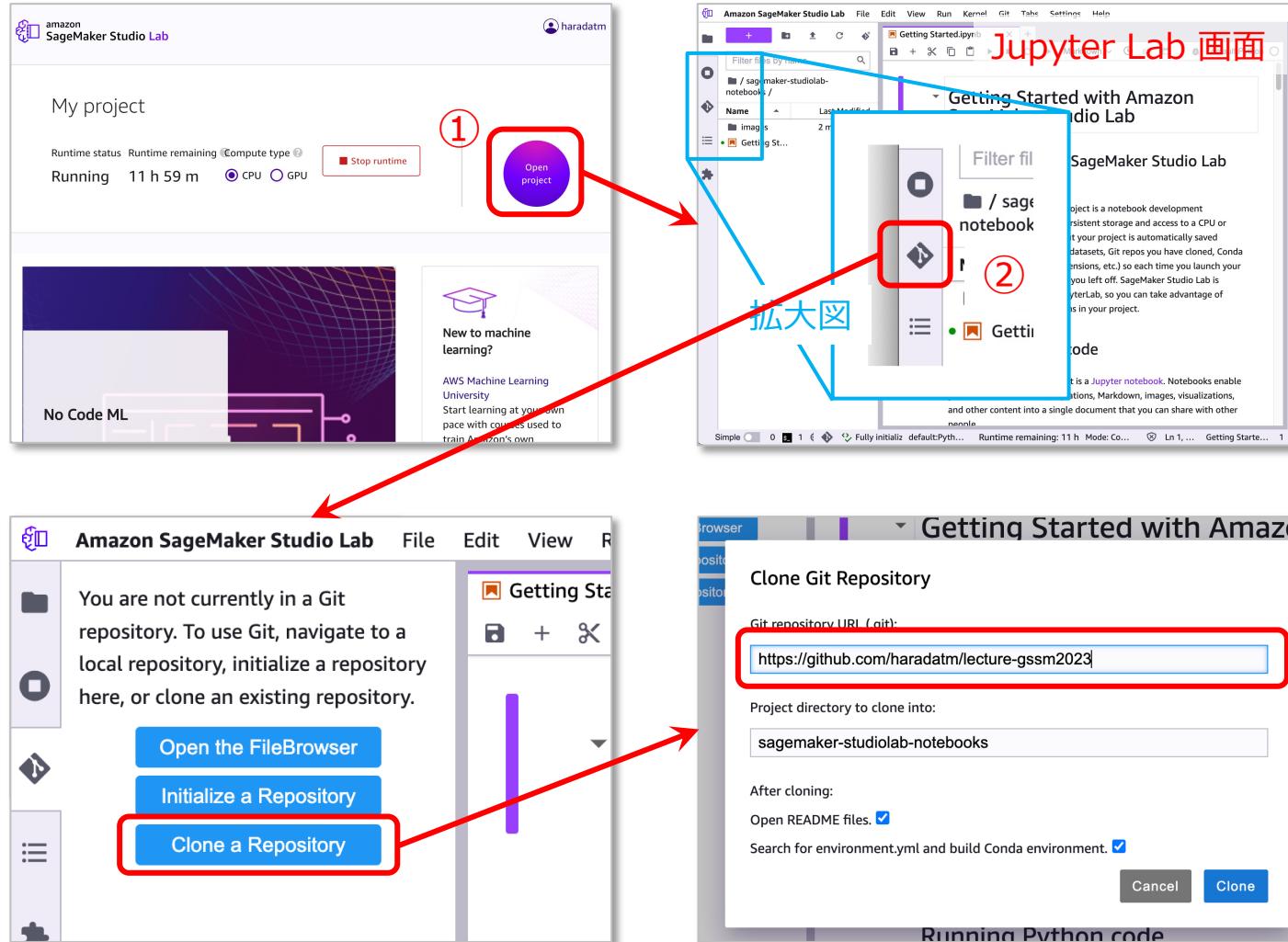
- ① 画面右上の「Sign in」ボタンを押す
- ② Eメールアドレス/ユーザー名、パスワードを入力する
- ③ 「Sign in」を押してプロジェクトのページを開く

2. Runtime を起動する

- ④ 「My Project」の「Select compute type」から「CPU」を選択する
- ⑤ 「Start runtime」を押す
- ⑥ 起動時に多要素認証を求められた場合、使用可能なデバイスで認証する

演習環境の準備

● Jupyter Lab を起動して、教材を開いてください



3. Jupyter Lab を起動する

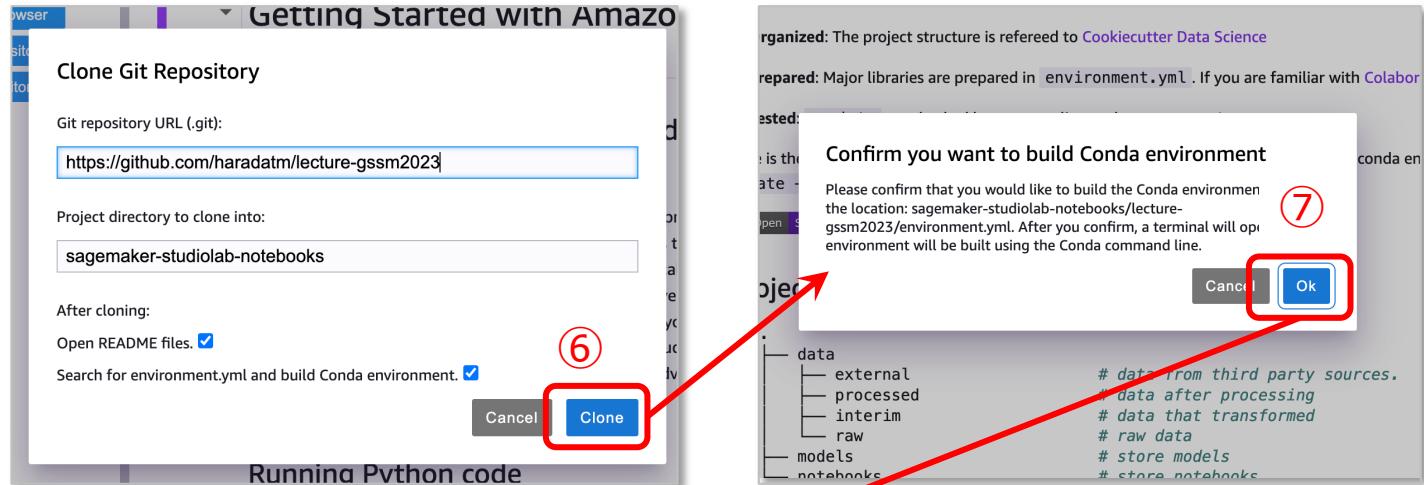
- ① ランタイムが開始したら「Open project」を押す

4. 教材を開く

- ② 「Git」  ボタンを押す
- ③ 「Clone a Repository」を押す
- ④ 「Git repository URL ...」に
<https://github.com/haradatm/lecture-gssm2023> を入力する
- ⑤ (任意) 「Project directory to ...」に保存先のパスを入力する
例) 「sagemaker-studiolab-notebooks」

演習環境の準備

● Jupyter Lab を起動して、教材を開いてください（続き）



4. 教材を開く（続き）

⑥ 画面右下の「Clone」を押す

⑦ ポップアップした画面で「OK」を押す

この後、処理完了まで**7分程度**待ちます

⑧ 「done」が表示される

⑨ 続いて「プロンプト」が表示されれば完了

The screenshot shows a terminal window with the following output:

```
Building wheel for jpaniize-matplotlib (setup.py): started
Building wheel for jpaniize-matplotlib (setup.py): finished with status 'done'
Created wheel for jpaniize-matplotlib: filename=jpaniize_matplotlib-1.1.3-py3-none-any.whl size=4120258 sha256=c789fb6583b8222bafae8c92250171d679d0af6263ac8632b174995a6d5fb1
Stored in directory: /home/studio-lab-user/.cache/pip/wheels/da/a1/71/b8faeb93276fed10edffcc20746f1ef6f8d9e071eee8425fc
Successfully built jpaniize-matplotlib
Uninstallable collected packages: pluggy,ini configparser,coverage,pytest,cov-jpaniize-matplotlib
Successfully installed coverage-7.2.7 ini configparser-2.0.0 jpaniize-matplotlib-1.1.3 pluggy-1.0.0 pytest-7.3.1 pytest-cov-4.1.0
done
# To activate this environment, use
# $ conda activate gssm2023
#
# To deactivate an active environment, use
# $ conda deactivate
```

A blue box labeled '拡大図' (Zoomed-in view) surrounds the 'done' message and the terminal prompt. Red boxes labeled ⑧ and ⑨ point to these respective areas.

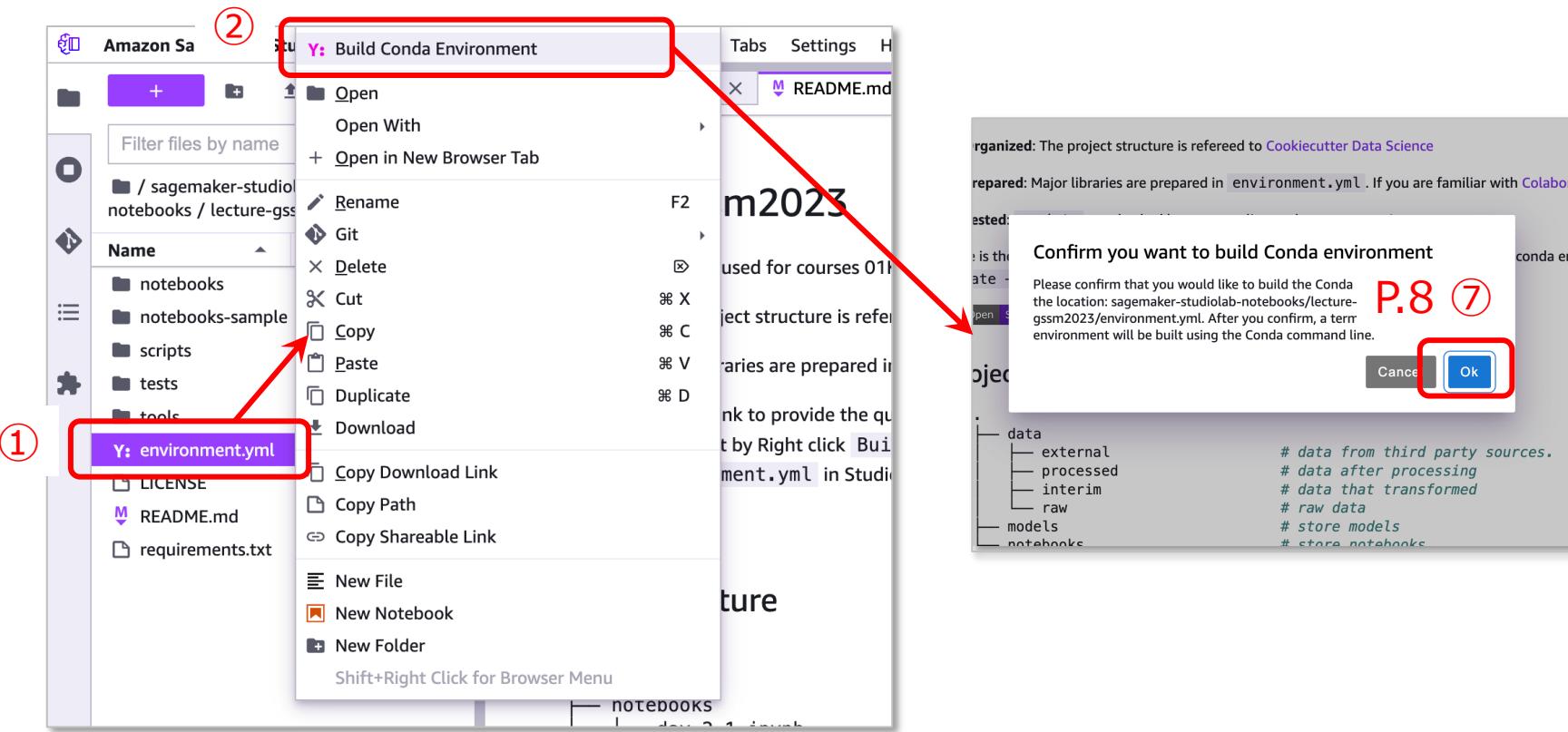
注意：

⑦で「OK」を押さなかった場合、
⑧や⑨の表示がされない場合は
後述の「トラブルシューティング」
から再開してください

(参考) ブラウザで「environment.yml」を選択する

- P.8 の⑦で「OK」を押さなかった場合や、⑧や⑨の表示がされない場合

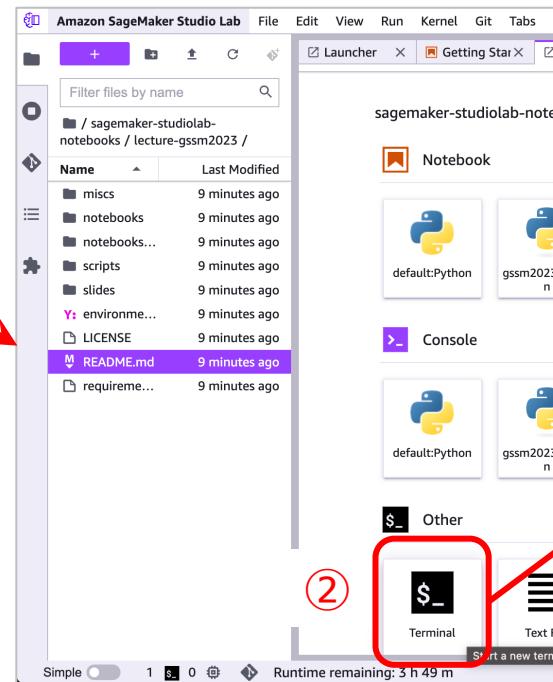
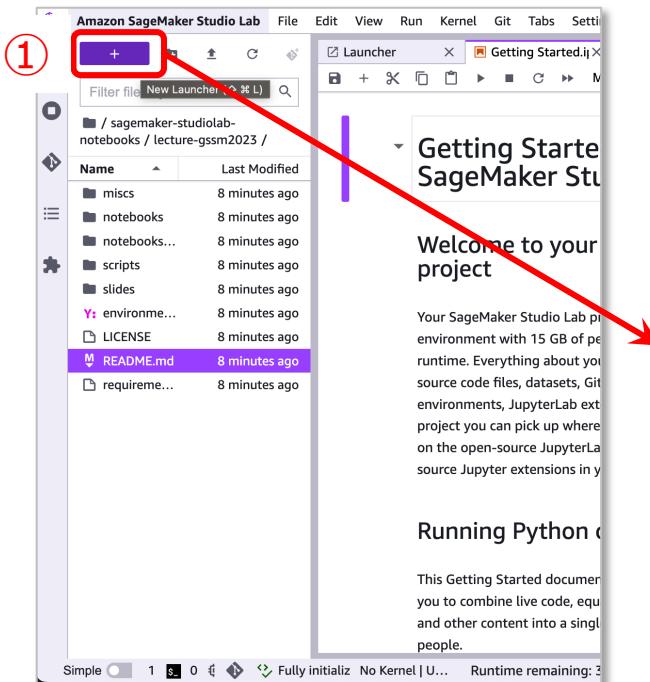
- ① 画面左の **File Browser** から「**environment.yml**」を選択する
- ② 右クリックメニューから「**Build Conda Environment**」を選択する → P.8 ⑦ へ



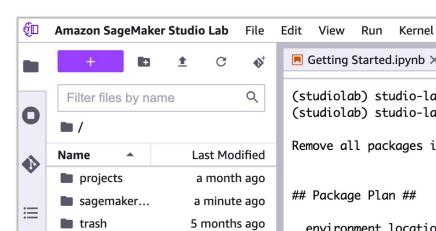
(参考) ブラウザによる開発環境構築

● (いちからやり直すために) プロジェクトを完全削除する **注意: この操作は慎重に行ってください**

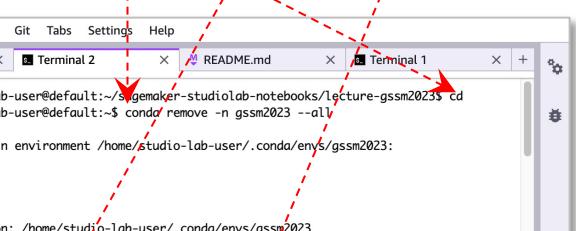
- ① 画面上の **+** ボタンで **Launcher** を開く
- ② 画面上の **Launcher** から **Terminal** を開く
- ③ **Terminal** で右のコマンドを実行する
- ④ **Stop runtime** で Runtime を停止 → P.7 へ



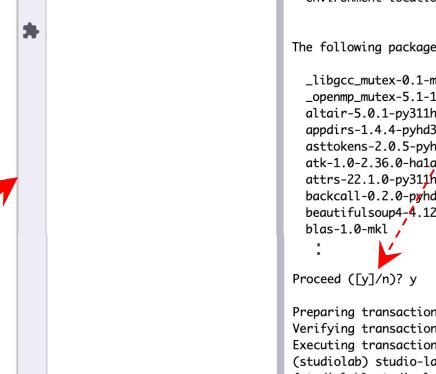
```
# ホームディレクトリへ移動  
cd  
# 仮想環境を削除  
conda remove -n gssm2023 --all # Proceed ([y]/n)? Y  
# プロジェクトを削除  
rm -fr sagemaker-studiolab-notebooks/lecture-gssm2023
```



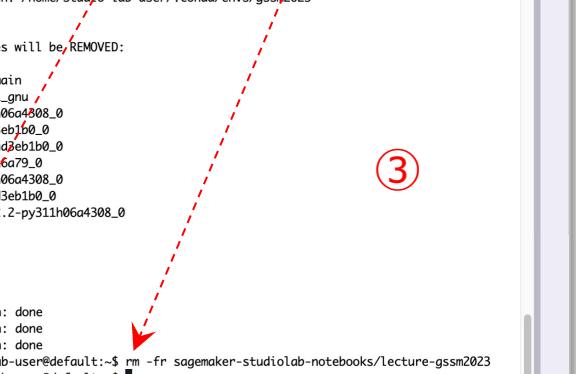
cd



Remove all packages in environment /home/studio-lab-user/.conda/envs/gssm2023:
Package Plan ##
environment location: /home/studio-lab-user/.conda/envs/gssm2023



The following packages will be REMOVED:
libgcc_mutex-0.1-main
_openmp_mutex-5.1-1_gnu
altair-5.0.1-py311h06a4308_0
appdirs-1.4.4-pyh3eb1b0_0
asttokens-2.0.5-pyhd3eb1b0_0
atk-1.0-2.36.0-ha0a79_0
attr-22.1.0-py311h06a4308_0
backcall-0.2.0-pyd3eb1b0_0
beautifulsoup4-4.12.2-py311h06a4308_0
blas-1.0-mkl
:
Proceed ([y]/n)? y



Preparing transaction: done
Verifying transaction: done
Executing transaction: done
(studio-lab) studio-lab-user@default:~\$ rm -fr sagemaker-studiolab-notebooks/lecture-gssm2023
(studio-lab) studio-lab-user@default:~\$

テキストマイニングの手順

テキストマイニング

- 驚異的な大量の文書データに記述されている多種多様な内容を対象として、その相関関係や出現傾向などから新たな知識を発見する [那須川,1999]
- 市場調査や販売戦略の立案、製品やサービス改善、顧客対応の改善に役立てたい
 - アンケート、レビューサイトのクチコミ、ツイートなど
- 最近では、報道番組などで Twitter 分析を取り上げることも多い
 - 震災、選挙、新型コロナウィルスなど

クチコミ分析の例 — コックroach

- パッケージ描かれたイラストが嫌 → 変更後、前年比2倍の出荷



出典: http://www.kincho.co.jp/seihin/insecticide/go_aerosol/gokiburi_u_spray/index.html

クチコミ分析の例 — 都市観光ホテル

- 温泉街の集客低下

→ 浅間温泉の観光客は松本市内に宿泊



星野リゾート
ホームページより

- 全国 の都市部にあるビジネスホテルを調査

→ 宿泊客の 6割 はビジネス客でなく「観光客」

→ 一方で、料金に不満はないものの旅のテンションが下がってしまう



- 都市型ホテルがどうか変われるか → 都市観光ホテル

引用: <https://travel.watch.impress.co.jp/docs/news/1056715.html>
<https://www.hoshinoresorts.com/brand/omo/>

クチコミ分析の例 — ???

- クチコミ分析の例をあげてください

クチコミサイトの例 — 楽天トラベル

● ホテルのクチコミ数: 1,325万件 ※年間約60~80万

The screenshot shows the Rakuten Travel website at <https://travel.rakuten.co.jp/review/>. The main heading is 'お客様の声' (Customer Reviews) with the number '13,246,463' in red. Below it is a search bar for reviews and filters for domestic and overseas stays. A sidebar on the left lists new reviews, and a right sidebar highlights customer feedback.

新着！最新のクチコミ

2023年5月 27日 更新

国内宿泊 海外ホテル

2023-05-26 23:58:52 宝泉寺温泉 ペットと泊まれる宿 季の郷 山の湯のクチコミ (157件) ★★★★★ 4.63

2023-05-26 23:54:17 三喜旅館のクチコミ (18件) ★★★★★ 5

「お客様の声」には、実際にご利用になった方のご意見・ご感想が満載です。

国内宿泊 高速バス
ペットホテル 海外ホテル

経年変化:

780万件 (2015)
→ 836万件 (2016)
→ 900万件 (2017)
→ 973万件 (2018)
→ 1,042万件 (2019)
→ 1,098万件 (2020)
→ 1,165万件 (2021)
→ 1,237万件 (2022)
→ 1,325万件 (今回)
※ 2021/5/27現在

R 鶴川シーワールドホテル クチ ×

HARADA Tomohiko

travel.rakuten.co.jp/HOTEL/2910/review.html

G

楽天
トラベル 宿・航空券・ツアー予約

楽天カード入会で2,000ポイントプレゼント カード GORA 楽天市場
 楽天トラベルの使い方 サイトマップ ヘルプ Languages -
 ようこそ、楽天トラベルへ 会員登録 ログイン 予約の確認・キャンセル

楽天
スーパーDEAL
 30%以上ポイントバックも!

国内旅行 国内ツアー レンタカー 高速バス 海外旅行 海外ツアー 海外航空券 海外ホテル 割引クーポン 懸賞広場 観光案内

楽天トラベルトップ > 全国 > 千葉県 > 外房（鶴川・勝浦・御宿・茂原）> 鶴川温泉 > 鶴川シーワールドホテル クチコミ・感想・情報

鶴川シーワールドホテル

★★★★★ 4.12 クチコミ・お客様の声(886件) この宿泊施設をお気に入りに追加 メルマガ 幹事さん機能
[友達にメール](#) [シェアする](#) [3](#) [Twitter](#) [Facebook](#) [Google+](#) [Pinterest](#)

日程からプランを探す

国内宿泊
 ANA 航空券+宿泊
 JAL 航空券+宿泊
 日帰り・デイユース
 日付未定

チェックイン

2015/06/21

チェックアウト

2015/06/22

ご利用部屋数

1 部屋

ご利用人数

1部屋目 :

大人 (1 人) 子供 (0 人)
金額(1部屋1泊あたり消費税込)

下限 [制限なし] 上限 [制限なし]

検索

最近見た宿泊施設

11軒の閲覧履歴があります
[ページ1/6]

もっと見る

施設開通情報

◎鶴川シーワールドホテル

★トップページ★

◎鶴川シーワールドホテル

★構造ニュース★

施設開通情報

<div data-bbox="22

R 鴨川シーワールドホテル クラ ×

HARADA Tomohiko

travel.rakuten.co.jp/HOTEL/2910/review.html

Q ☆ G

★お部屋★

- 鴨川シーワールドホテル
- レストラン★
- 鴨川シーワールドホテル
- 温泉大浴場★
- 鴨川シーワールドホテル
- 館内施設★
- 鴨川シーワールドホテル
- ★よくあるご質問★
- 鴨川シーワールドホテル
- アクセス★
- 鴨川シーワールドホテル設備・アメニティ・基本情報
- 鴨川シーワールドホテル写真・画像
- 鴨川シーワールドホテル地図・アクセス
- 鴨川シーワールドホテルチケット
- 鴨川シーワールドホテル温泉

外國語サイト

- Book Kamogawa Sea World Hotel
- Hotels in Sotobo(Kamogawa, Katsuura, Onjuku, Mabora)
- KAMOGAWA SEA WORLD HOTEL 預訂
- 外房(鴨川)・勝浦・御宿・茂原)酒店一覧

★★★★★ 2

投稿者さんの 鴨川シーワールドホテル のクチコミ (感想・情報)

投稿者さん

2015年06月11日 17:03:57

良かったところ

- ・部屋からの景色（朝日最高でした）
- ・食事（品数が多く、朝夕とも良かったです）
- ・フロントの方の対応（お姉さんがとても頑張っていました）以上。

掃除が行き届いているとの口コミを多く見ましたが、そろは思いませんでした。

気にかかることは多々ありました、フロントのお姉さんが一生懸命で、その笑顔に救われた思いです。

レビューを評価して不適切なレビューを報告する
このレビューは参考になりましたか?

いいえ

いいえ

旅行の目的 … レジャー

同伴者 … 家族

宿泊年月 … 2015年06月

ご利用の宿泊 【洋室 禁煙・特別室】 お部屋からシャチやイルカも見える シーワールドと海一望宿泊プラン

ご利用のお部屋 【wa5シーワールド】が見える特別室禁煙【洋室】】

★★★★★ 4

投稿者さんの 鴨川シーワールドホテル のクチコミ (感想・情報)

投稿者さん

2015年06月11日 07:33:49

夫、2歳半と5ヶ月の子どもの4人で宿泊しました。

【立地】当たり前ですが鴨川シーワールドにとても近く、ゆっくり館内を見学できました。

【部屋】至って普通です。(古いからか、勝の声は少し聞こえます。)トレイ掃除などはしっかりされていました。清淨機などもTEL一本ですぐに届けて下さいました。

【食事】夜朝共にバイキング。イスですが子ども用イス、エプロン、ベビーベッドを用意して下さっています。キッズスペースも食事時間中に専門のスタッフの方がおりゆっくり食事ができました。

【風呂】小さな子ども(赤ちゃん)用のグッズ(ペリーベッド、コーナー、バス、おもちゃ、泡ソーブ、え vess のアイス)が揃っていました。お子さん連れも多く、気兼ねなく楽しめました。しかしお風呂がひとつしかないため、温泉を楽しむという雰囲気ではなく、錢湯の湯湯が温泉という感じです。

また、2・3時頃にお風呂に行くと、アメニティやシャンプーが空だったのは少し残念でした。

【サービス】受付スタッフの皆さんとても親切、丁寧です。チェックアウト後に子どもの薬を冷蔵庫にいておいて欲しいとダメ元で頼みとぐく入

鴨川シーワールドホテル

2015年06月11日 19:25:50

この度は、ご利用頂きまして誠にありがとうございました。

客室内清掃の件、大変申し訳ございませんでした。

重要改善として、早急に対応いたします。
今後は、この様な事の無いよう、清掃・点検を強化いたします。

フロントスタッフへのお言葉、
誠にありがとうございます。

モチベーションアップに繋がりますので、
お客様からの声として、
スタッフと共に共有させて頂きます。

機会がございましたら、またご利用をお待ちしております。

いい値バリュープラン
【最安料金 (国安)】10,186円~
(消費税込11,000円~)

【当日15:50からアシカと記念写真】笑うアシカと一緒にアフリカパン 室数限定
【最安料金 (国安)】10,278円~
(消費税込11,100円~)

【当日13:40~エコーアクアームズミュニケーションタイム】1日3組限定
【最安料金 (国安)】10,278円~
(消費税込11,100円~)

【夜の水族館探検付】3ヶ月~10月の火・木曜日限定プラン
【最安料金 (国安)】10,278円~
(消費税込11,100円~)

【当日14:50からイルカと一緒にパリティ】2室限定
【最安料金 (国安)】10,463円~
(消費税込11,300円~)

今しかない!アワビ料
理付サシワールド入園
バス料付で大満足
5月・6月の木~木曜日
限定プラン
【最安料金 (国安)】10,926円~
(消費税込11,800円~)

【便利な赤ちゃんグッズ付】
初お泊りお母さんも嬉しい★赤ちゃん
うつ寝プラン
【最安料金 (国安)】10,926円~
(消費税込11,800円~)

お子様にも大好評!オーサンシャンプーブラン
【最安料金 (国安)】11,112円~
(消費税込12,000円~)

【80cmのジャンボボイサイズ】
海の王者シャチぬいぐるみプラン
【最安料金 (国安)】11,204円~
(消費税込12,100円~)

房総2大テーマパーク満喫「マザーランドチケット」付プラン
【最安料金 (国安)】11,389円~
(消費税込12,300円~)

【当日14:50~イルカ

鴨川シーワールドホテルのクチコミ・お客様の声

[●ホテル・旅行のクチコミTOPへ](#)

総合評価

4.12

アンケート件数：886件

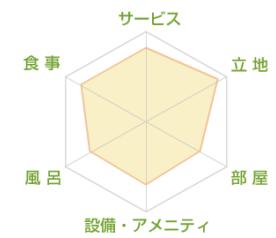
評価内訳

- 5点
- 4点
- 3点
- 2点
- 1点

236件
302件
47件
15件
9件

項目別の評価

| | |
|----------|------|
| サービス | 4.11 |
| 立地 | 4.61 |
| 部屋 | 3.53 |
| 設備・アメニティ | 3.62 |
| 風呂 | 3.53 |
| 食事 | 4.10 |



総合 2

投稿者さんの 鴨川シーワールドホテル のクチコミ (感想・情報)



投稿者さん

2015年06月11日 17:03:57

良かったところ

- ・部屋からの景色（朝日最高でした）
- ・食事（品数が多く、朝夕とも良かったです）
- ・フロントの方の対応（お姉さんがとても頑張っていました）以上。

掃除が行き届いているとの口コミを多く見ましたが、それは思いませんでした。

気にかかることは多々ありましたが、フロントのお姉さんが一生懸命で、その笑顔に救われた思います。

評価

... 総合 2

- | | |
|----------|---|
| サービス | 2 |
| 立地 | 4 |
| 部屋 | 4 |
| 設備・アメニティ | 2 |
| 風呂 | 2 |
| 食事 | 4 |

旅行の目的

... レジャー

同伴者

... 家族

宿泊年月

... 2015年06月



鴨川シーワールドホテル

2015年06月11日 19:32:50

この度は、ご利用頂きまして誠にありがとうございます。

客室内清掃の件、大変申し訳

重要改善として、早急に対応いたします。

今後は、この様な事の無いように、清掃・点検を強化いたします。

テキストデータ

フロントスタッフへのお言葉
誠にありがとうございます。

セラピーアップに繋がる
お客様からの声として、
スタッフと共有させて頂きます。

数値評価

機会がございましたら、またご利用をお待ちしております。

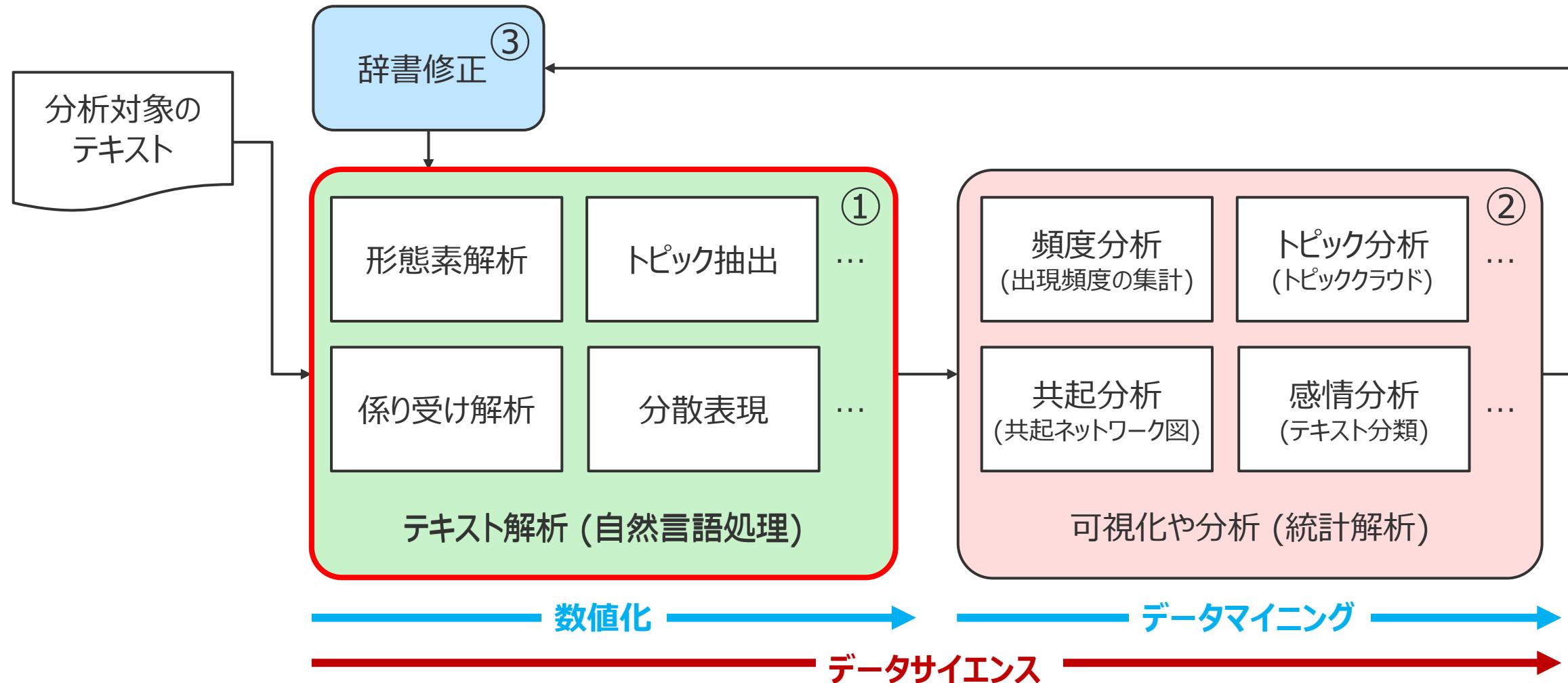
テキストマイニングの手順

- データをよく知る
 - データ件数や構成比を集計 → データを理解する
 - 旅行目的別の人気エリアは?
 - 同伴者別の人気エリアは?
 - 数値評価による人気エリアの差異は?
- テーマを設定する
 - 解決すべき課題を決める → 分析目的を明確にする
 - 数値評価が低い原因是?
 - 高評価の施設に学ぶ改善点は?
- テキスト分析に取り組む
 - これら課題を解決するために、テキスト分析を実施

テキスト解析 (1)

テキスト分析の手順

①自然言語処理によりテキストを数値化する → ②統計解析や可視化を行う → ③結果を読み解きながら解析のための辞書を編纂する → 分析のサイクルを回していく(①へ)



代表的なテキスト解析器

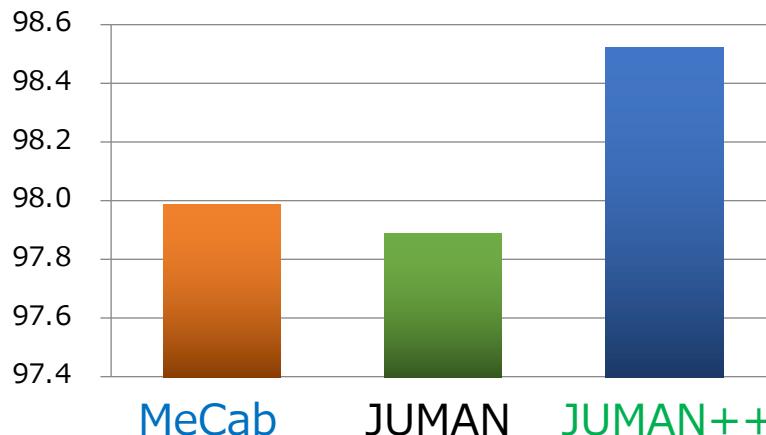
- 速度重視では **MeCab**、精度と出力情報の豊富さ重視では **JUMAN++** がお勧め

出典: <https://taku910.github.io/mecab/> をもとに加筆修正

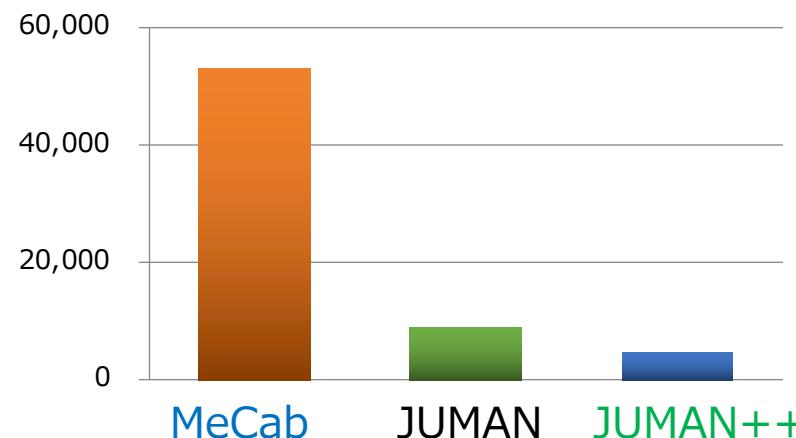
| 形態素解析器 | ChaSen | MeCab | JUMAN | JUMAN++ |
|--------|----------------------|---------|-------|---------|
| コスト推定 | HMM | CRF | 人手 | RNNLM |
| 探索方法 | 接続コスト最小法 (ビタビアルゴリズム) | | | |
| 係り受け解析 | Cabocha | CaboCha | | KNP |

JUMAN++ 深層学習を使った手法で、自然な言葉の繋がりを考慮した

単語分割+品詞タグ付け精度 (F1)



処理速度 (文/秒)



学習・評価データ

京都大学テキストコーパス (NEWS),
京都大学ウェブ文書リードコーパス (WEB)

RNN言語モデルの学習

Webコーパス 1000万文

出所:

https://drive.google.com/file/d/1DVnrsWw4skRqC8jU6_RkeofOQEHFwctcview?usp=sharing

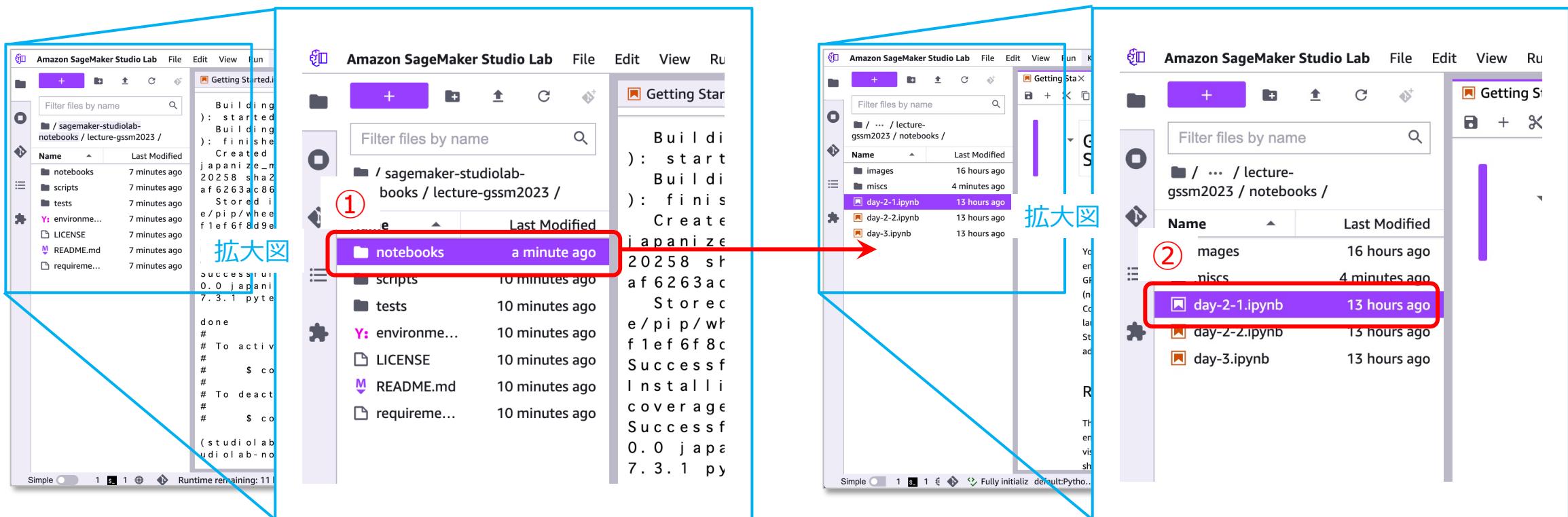
- Megagon Labs と国立国語研究所の共同研究結果として公開された OSS の日本語自然言語処理ライブラリ
 - ・「著作権表示」と「MIT ライセンスの全文」を記載する、という2条件のみで商用利用が可能
- **spaCy** (機械学習を組み込んだ自然言語処理ライブラリ) 上で動作するので、係り受け解析や固有表現抽出などの機能も利用可能
 - ・ 形態素解析には **Sudachi** (徳島人工知能NLP研究所)を利用、辞書は半年に約1回の頻度で更新されている (らしい)
 - ・ 20億文以上のWebテキストで事前学習した **Transformers** モデルも利用可能

ただし、高い利便性や機能の一方で処理が遅い (形態素解析でMeCabの10倍ぐらい)

演習 — Jupyter ノートブックの使い方

● day-2-1.ipynb を開いてください

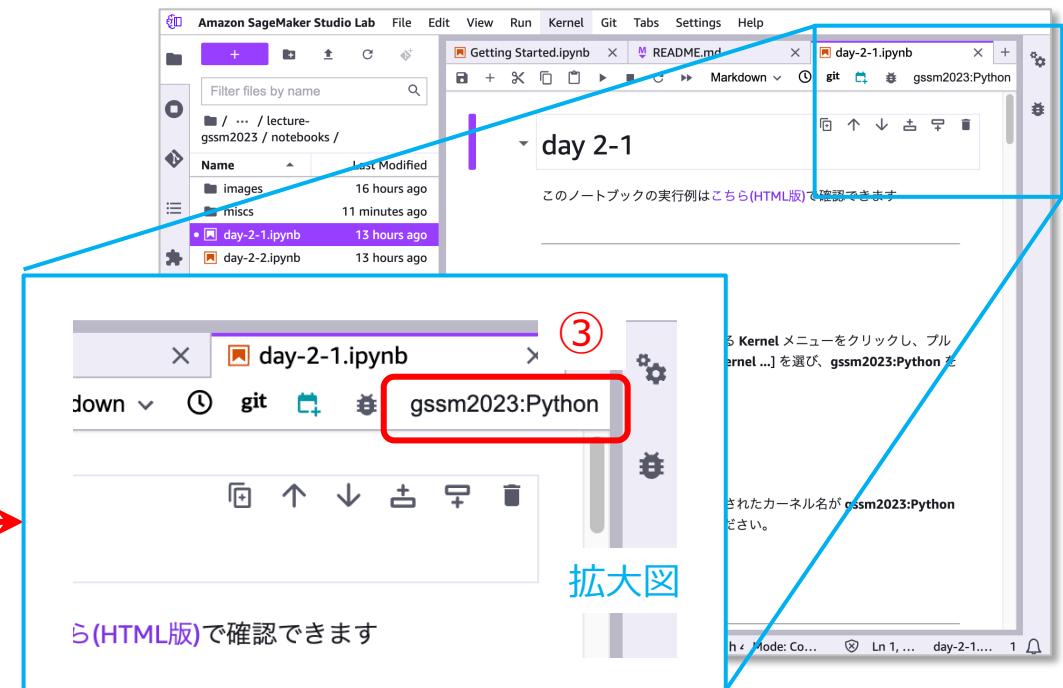
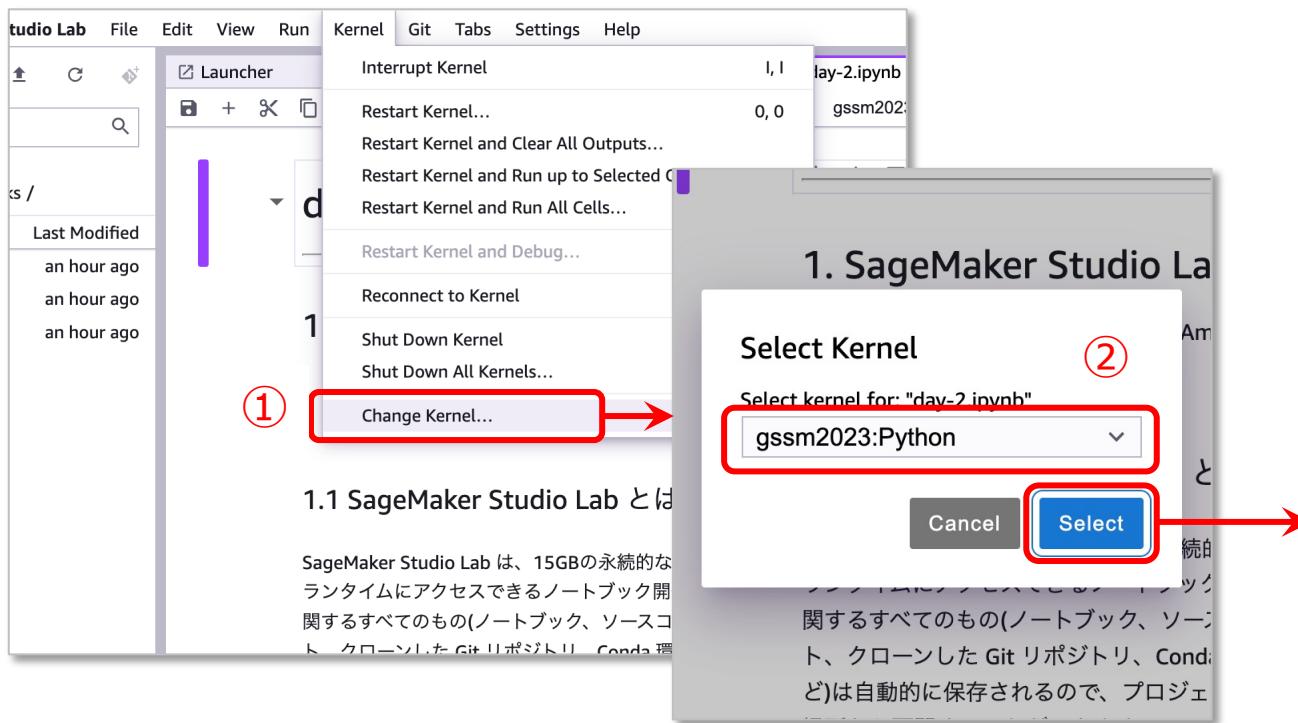
- ① 画面左の **File Browser** から ① **notebooks** をフォルダを開く
- ② さらに **day-2-1.ipynb** ノートブックを開く



演習 — Jupyter ノートブックの使い方

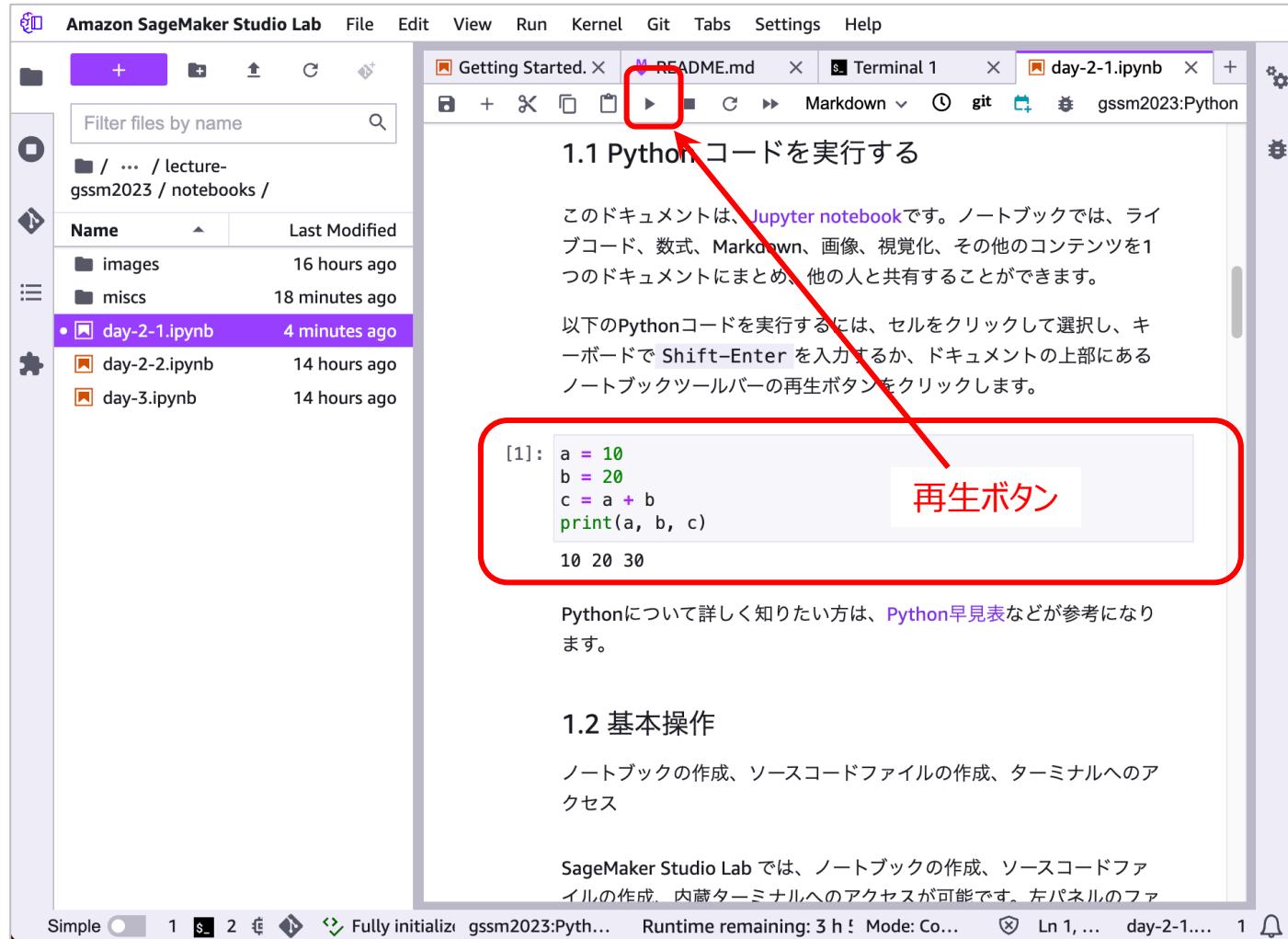
● カーネル gssm2023:Python を選択してください !重要!

- ① ページ上部の **Kernel** メニューから「Change Kernel ...」を選ぶ
- ② ポップアップ画面から「gssm2023:Python」を選択し、「Select」を押す
- ③ 右上隅にカーネル名「gssm2023:Python」が表示されていることを確認する



演習 — Jupyter ノートブックの使い方

● Python コードを実行する



- Jupyter ノートブックでは、Python のプログラムを実行することができます
(別のパッケージをインストールすることで、R 言語や C++ も実行できます)
- ノートブック上で、Pythonコードを実行するには、**セルをクリックして選択し、キーボードで Shift-Enter を入力するか、ドキュメントの上部にあるノートブックツールバーの再生ボタンをクリックします**

練習：左図の赤枠にあるセルをクリックし、セル内の計算を実行する

演習 — テキスト解析 (1)

● 形態素解析器 MeCab と、係り受け解析器 CaboCha をインストールする

The screenshot shows a Jupyter notebook titled "day-2-1.ipynb" in the "Getting Started" tab of Amazon SageMaker Studio Lab. The notebook contains five numbered steps:

- ① 2.1 MeCab のインストール (目安:約3分)
[4]:
!bash .. /scripts/install_mecab.sh > install_mecab.log 2>&1
!tail -n 1 install_mecab.log
Successfully installed mecab-python-0.996
- ② Successfully installed mecab-python-0.996 と表示されれば、インストール成功です。
- ③ 2.2 CaboCha のインストール (目安:約5分)
[5]:
!bash .. /scripts/install_cabocha.sh > install_cabocha.log 2>
!tail -n 1 install_cabocha.log
Successfully installed cabocha-python-0.69
- ④ Successfully installed cabocha-python-0.69 と表示されれば、インストール成功です。
- ⑤ 2.3 Kernel のリスタート
ページ上部のメニューにある **Kernel** メニューをクリックし、プルダウンメニューから [Restart Kernel ...] を選択してください。

A red box highlights the first step, and a red arrow points to the play button icon in the toolbar above the code cell, labeled "再生ボタン".

「2.1 MeCab のインストール」

- ① セルをクリックして選択し、再生ボタンを押す
この後、処理完了まで約3分程度待ちます
- ② 「Successfully ...」を確認する

「2.2 CaboCha のインストール」

- ③ セルをクリックして選択し、再生ボタンを押す
この後、処理完了まで約5分程度待ちます
- ④ 「Successfully ...」を確認する

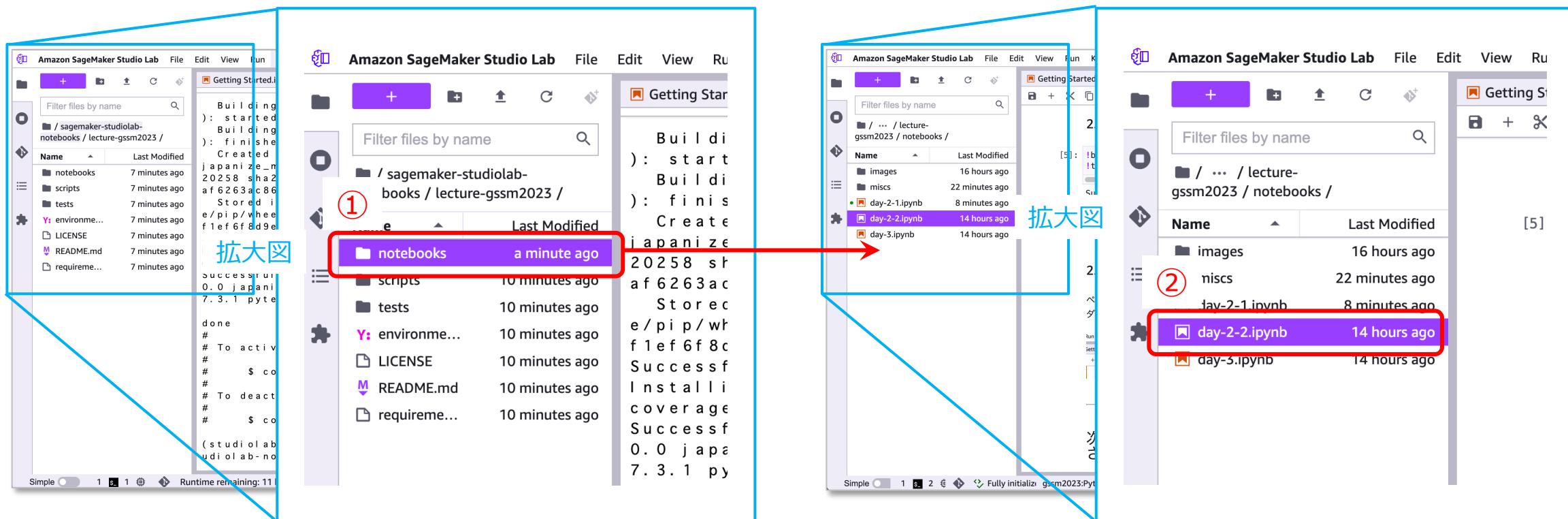
「2.3 Kernel のリスタート」

- ⑤ メニューバーにある **Kernel** メニューをクリックし、プルダウンメニューから [Restart Kernel ...] を選択する

演習 — テキスト解析 (1)

● day-2-2.ipynb を開いてください

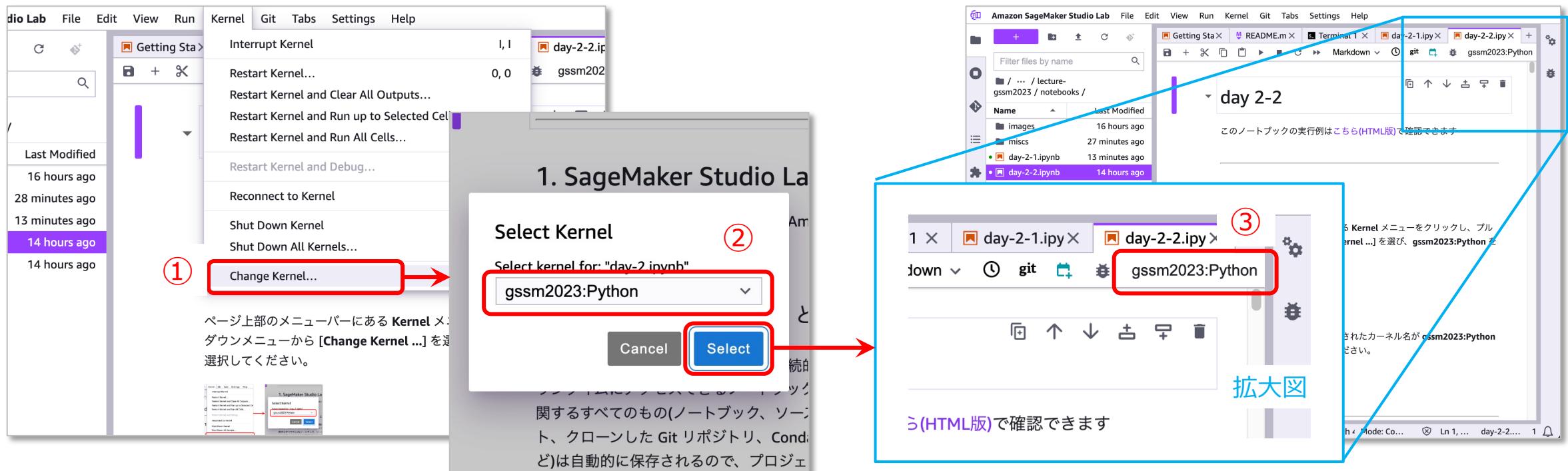
- ① 画面左の File Browser から ① notebooks をフォルダを開く (既に開いている場合はスキップ)
- ② 次に day-2-2.ipynb ノートブックを開く



演習 — テキスト解析 (1)

● カーネル gssm2023:Python を選択してください !重要!

- ① ページ上部の **Kernel** メニューから「Change Kernel ...」を選ぶ
- ② ポップアップ画面から「gssm2023:Python」を選択し、「Select」を押す
- ③ 右上隅にカーネル名「gssm2023:Python」が表示されていることを確認する



演習 — テキスト解析 (1)

● 形態素解析を行う (コマンドライン実行と同じ形式)

3.1 MeCab を使う

(1) そのまま出力してみる

①

```
import MeCab

tagger = MeCab.Tagger("-r ..//tools/usr/local/etc/mecabrc")
print(tagger.parse("今日はいい天気です"))
```

```
今日    名詞,副詞可能,*,*,*,*,今日,キヨウ,キヨー
は      助詞,係助詞,*,*,*,*,は,ハ,ワ
いい   形容詞,自立,*,*,形容詞・イイ,基本形,いい,イイ,イイ
天気   名詞,一般,*,*,*,*,天気,テンキ,テンキ
です   助動詞,*,*,*,特殊・デス,基本形,です,デス,デス
EOS
```

- ① セルをクリックして選択し、再生ボタンを押す
 - この方法では、コマンドライン実行した場合と同じ形式で出力されます
 - ただし、テキスト解析では、**テキストを数値化し、統計処理を行う必要**があります
 - そこで、**統計処理で扱いやすい DataFrame 型**(テーブル形式)に格納します → 次ページ

演習 — テキスト解析 (1)

● 形態素解析を行う (DataFrame 型に格納する)

②

```
import pandas as pd

node = tagger.parseToNode("今日はいい天気です")
features = []
while node:
    features.append(node.feature.split(','))
    node = node.next

columns = [
    "品詞", "品詞細分類1", "品詞細分類2", "品詞細分類3", "活用型", "活用形", "基本形",
    "読み", "発音",
]
pd.DataFrame(features, columns=columns)
```

[2]:

| | 品詞 | 品詞細分類1 | 品詞細分類2 | 品詞細分類3 | 活用型 | 活用形 | 基本形 | 読み | 発音 |
|---|---------|--------|--------|--------|--------|-----|-----|-----|-----|
| 0 | BOS/EOS | * | * | * | * | * | * | * | * |
| 1 | 名詞 | 副詞可能 | * | * | * | * | 今日 | キョウ | キョー |
| 2 | 助詞 | 係助詞 | * | * | * | * | は | ハ | ワ |
| 3 | 形容詞 | 自立 | * | * | 形容詞・イイ | 基本形 | いい | イイ | イイ |
| 4 | 名詞 | 一般 | * | * | * | * | 天気 | テンキ | テンキ |
| 5 | 助動詞 | * | * | * | 特殊・デス | 基本形 | です | デス | デス |
| 6 | BOS/EOS | * | * | * | * | * | * | * | * |

② セルをクリックして選択し、再生ボタンを押す

- この方法では、形態素解析器の出力を統計処理で扱いやすい DataFrame 型 (テーブル形式) に格納しています

練習: 入力文「**今日はいい天気です**」の内容を変更して、形態素解析(②)を行った結果を確認してください

演習 — テキスト解析 (1)

● 係り受け解析を行う（コマンドライン実行と同じ形式）

4.1 CaboCha を使う

(1) そのまま出力してみる

①

```
import CaboCha

cp = CaboCha.Parser("-r ../tools/usr/local/etc/cabocharc")
tree = cp.parse("今日はいい天気です")
print(tree.toString(CaboCha.FORMAT_LATTICE))
```

```
* 0 2D 0/1 -1.041733
今日 名詞,副詞可能,*,*,*,*,-,今日,キヨウ,キヨー
は 助詞,係助詞,*,*,*,-,は,ハ,ワ
* 1 2D 0/0 -1.041733
いい 形容詞,自立,*,*,-,形容詞・イイ,基本形,いい,イイ,イイ
* 2 -1D 0/1 0.000000
天気 名詞,一般,*,*,*,-,天気,テンキ,テンキ
です 助動詞,*,*,-,特殊・デス,基本形,です,デス,デス
EOS
```

- ① セルをクリックして選択し、再生ボタンを押す
- この方法では、コマンドライン実行した場合と同じ形式で出力されます
 - ただし、**係り元**や**係り先**の関係を把握するには、この出力形式でも、表形式でも直感的ではありません
 - そこで、**係り受け関係を確認し易いツリー形式**で出力します → 次ページ

演習 — テキスト解析 (1)

● 係り受け解析を行う（係り受けペアを抽出する）

②

```
# 構文木(tree)からチャンクを取り出す
def get_chunks(tree):
    chunks = {}
    key = 0
    for i in range(tree.size()):
        tok = tree.token(i)
        if tok.chunk:
            chunks[key] = tok.chunk
            key += 1
    return chunks

# チャンク(chunk)から表層形を取り出す
def get_surface(chunk):
    surface = ''
    begin = chunk.begin
    end = chunk.end
    for i in range(begin, end):
        surface += chunk[i]
    return surface
```

← 繰り返し呼ばれる処理などをまとめて関数として定義したもの

③

```
tree = cp.parse("今日はいい天気です")
chunks = get_chunks(tree)

for from_chunk in chunks.values():
    if from_chunk.link < 0:
        continue
    to_chunk = chunks[from_chunk.link]

    from_surface = get_surface(from_chunk)
    to_surface = get_surface(to_chunk)

    print(from_surface, '→', to_surface)
```

今日は → 天気です
いい → 天気です

② セルをクリックして選択し、再生ボタンを押す（③より前に一度実行しておく）

③ セルをクリックして選択し、再生ボタンを押す

- この方法では、係り受け解析器の出力を**係り元**と**係り先**の**関係**を持つ単語のペアを抽出しています

練習：入力文「**今日はいい天気です**」の内容を変更して、係り受け解析(③のみ)を行った結果を確認してください

実習用データについて

実習用のデータ (Webサイトクローリング)

● 楽天トラベル のクチコミデータ

- 収集期間は 2019-2020 および 2022-2023(～GW明け) の 2セット
- 以下の 10 エリアごと同数に 1,000件ずつ ランダムサンプリング
- データ件数は 1万件 × 2セット

| | | | |
|------|------|-----------------|-----------------------|
| レジャー | 5エリア | 登別、草津、箱根、道後、湯布院 | 1,000件 × 10エリア |
| ビジネス | 5エリア | 札幌、名古屋、東京、大阪、福岡 | = 計10,000件 |

実習用のデータ (Webサイトクローリング)

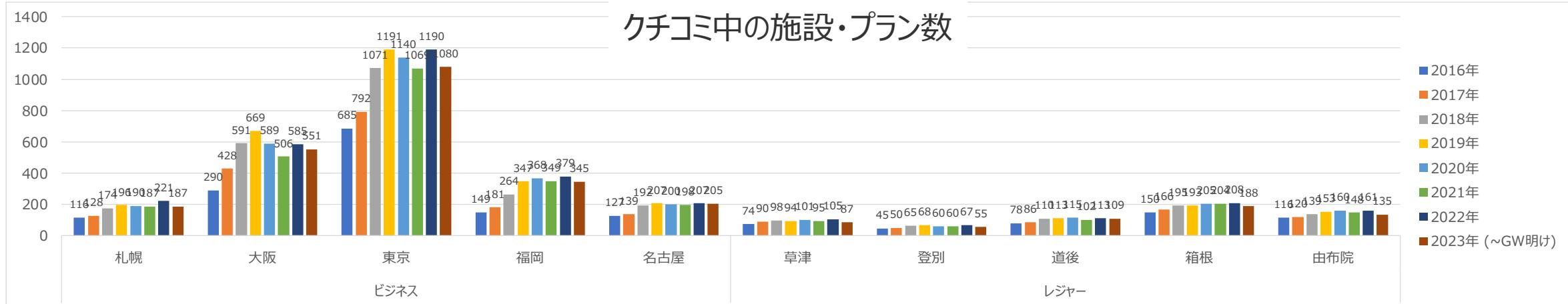
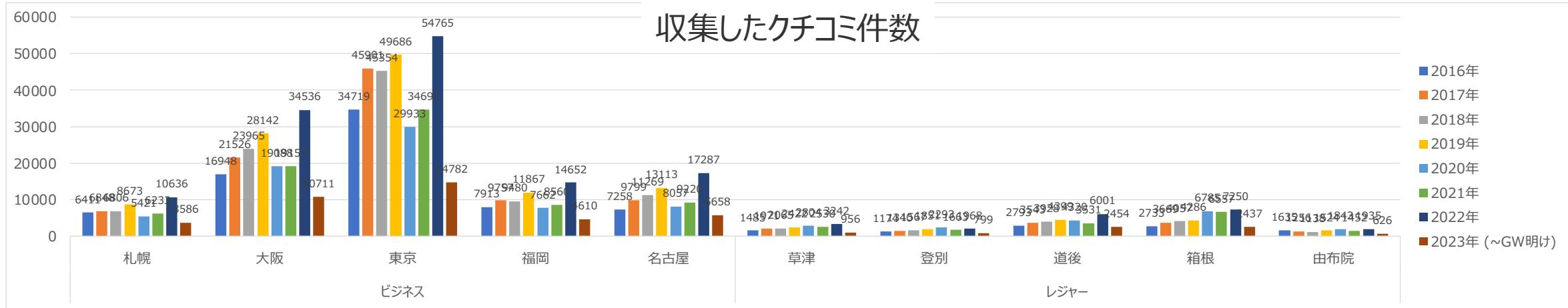
● 楽天トラベル のクチコミデータ

- 収集期間は 2019-2020 および 2022-2023(～GW明け) の 2セット
- 以下の 10 エリアごと同数に 1,000件ずつ ランダムサンプリング
- データ件数は 1万件 × 2セット
- データ項目は 18項目 (テキスト1項目+その他の属性17項目)

| | |
|---------------|--|
| 施設情報 | 4項目 カテゴリ, エリア, 施設番号, 施設名 |
| 口コミ | 1項目 コメント (テキスト) |
| ユーザー評価 | 7項目 総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事 |
| その他の分類 | 2項目 旅行の目的, 同伴者 |
| 宿泊日 | 1項目 宿泊年月 |
| ユーザー情報 | 3項目 ユーザー, 年代, 性別 |

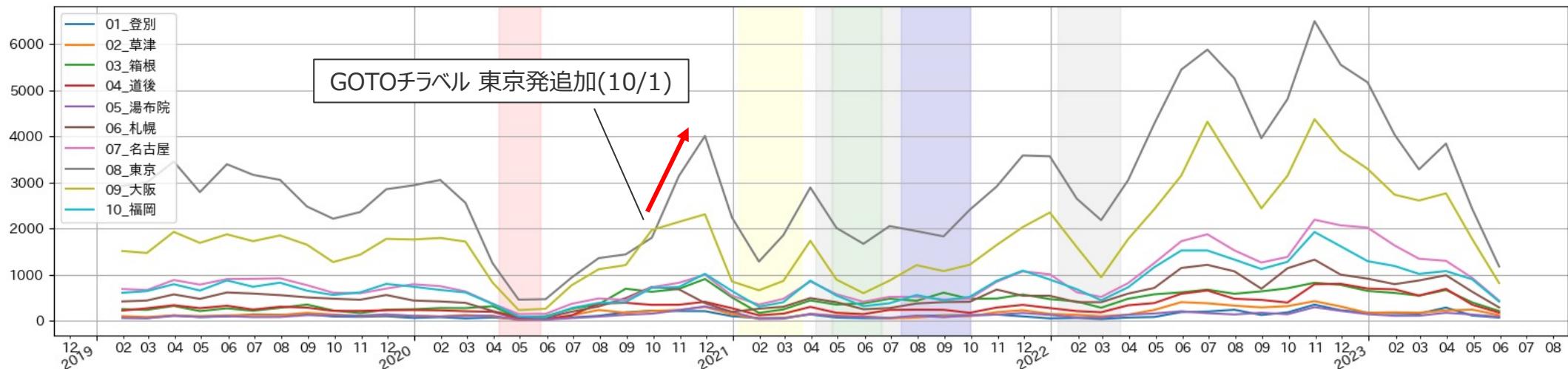
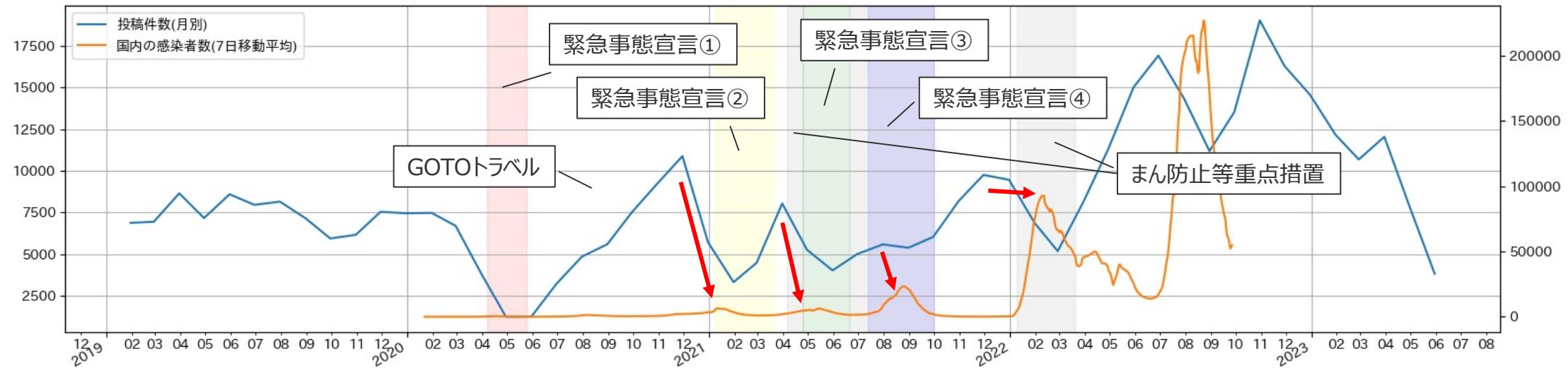
参考 — Webサイトクローリング

- 全量では 160.6万件、2020-2021は14.2万件、2022-2023は19.9万件



(参考) COVID-19 の影響

- クチコミの件数と感染者数の増減が連動 → クチコミ件数が一定の人流を反映している



実習用データ — ファイル一覧

● 実習用データは以下の通り → 主に「**rakuten-1000-2022-2023.xlsx**」を使用する

| ファイル名 | 件数 | データセット | 備考 |
|--|-----------|--|-------------------|
| <u>rakuten-1000-2022-2023.xlsx.zip</u> | 10,000 | <ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・エリアごと 1,000件 (ランダムサンプリング)・期間: 2022/1~2023 GW明け | 本講義の全体を通して使用する |
| <u>rakuten-1000-2020-2021.xlsx.zip</u> | 10,000 | <ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・エリアごと 1,000件 (ランダムサンプリング)・期間: 2020/1~2021/12 | 演習用 (年度で比較する場合など) |
| <u>rakuten-all-2022-2023-tsv.zip</u> | 142,061 | <ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・サンプリング前の全データ・期間: 2022/1~2023 GW明け | 参考用 |
| <u>rakuten-all-2020-2021-tsv.zip</u> | 198,885 | <ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・サンプリング前の全データ・期間: 2020/1~2021/12 | 参考用 |
| <u>rakuten-all-tsv.zip</u> | 1,659,396 | <ul style="list-style-type: none">・レジャー+ビジネスの 10エリア・サンプリング前の全データ・期間: 2009/3~2020/12 | 参考用 |

データ理解

テキストマイニングの手順

- データをよく知る
 - データ件数や構成比を集計 → データを理解する
 - 旅行目的別の人気エリアは?
 - 同伴者別の人気エリアは?
 - 数値評価による人気エリアの差異は?
- テーマを設定する
 - 解決すべき課題を決める → 分析目的を明確にする
 - 数値評価が低い原因是?
 - 高評価の施設に学ぶ改善点は?
- テキスト分析に取り組む
 - これら課題を解決するために、テキスト分析を実施

演習 — データ理解

● データをダウンロードする (ファイル名: rakuten-1000-2022-2023.xlsx)

The screenshot shows a Jupyter Notebook cell with the following content:

```
FILE_ID = "1n-uvGoH7XQhxexN57hYXuFrkGeHKp-HV"
!gdown --id {FILE_ID}
!unzip rakuten-1000-2022-2023.xlsx.zip
!ls -al rakuten-1000-2022-2023.xlsx
```

Output:

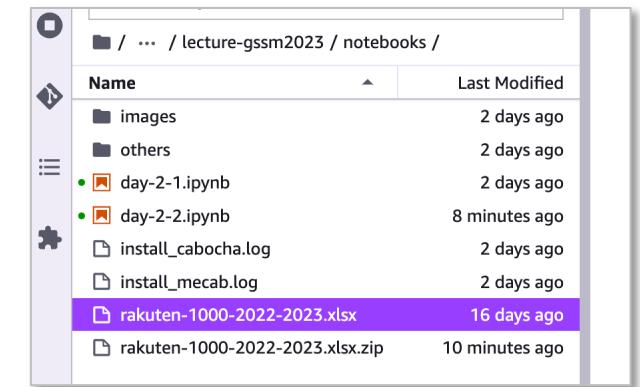
```
/home/studio-lab-user/.conda/envs/gssm2023/lib/python3.11/site-packages/gdown/cli.py:120: FutureWarning: Option '--id' was deprecated in version 4.3.1 and will be removed in 5.0. You don't need to pass it anymore to use a file ID.
  warnings.warn(
Downloading...
From: https://drive.google.com/uc?id=1n-uvGoH7XQhxexN57hYXuFrkGeHKp-HV
To: /home/studio-lab-user/sagemaker-studiolab-notebooks/lecture-gssm2023/notebooks/rakuten-1000-2022-2023.xlsx.zip
100%|██████████| 2.43M/2.43M [00:00<00:00, 21.3MB/s]
Archive: rakuten-1000-2022-2023.xlsx.zip
  inflating: rakuten-1000-2022-2023.xlsx
-rw-r--r-- 1 studio-lab-user users 2460755 May 28 02:22 rakuten-1000-2022-2023.xlsx
```

① A red box highlights the command line output, specifically the file download progress and final status line.

② Red arrows point from the number ② in the list below to the file name "rakuten-1000-2022-2023.xlsx" in the output and the file entry in the file browser on the right.

- ① セルをクリックして選択し、再生ボタンを押す
- ② ファイル **rakuten-1000-2022-2023.xlsx** が存在し、ファイルサイズが **2460755 byte** であることを確認する

※ ファイルは、**File Browser** に表示されます



演習 — データ理解

● データの読み込みと集計表の作成



4.2 データの読み込み (DataFrame型)

①

```
import pandas as pd  
  
df = pd.read_excel("rakuten-1000-2022-2023.xlsx")  
print(df.shape)  
display(df.head())  
  
(10000, 18)
```



4.3 集計

(1) エリア別の件数を表示する

②

```
display(df.pivot_table(index=['カテゴリー', 'エリア'], columns=None, values='コメント', aggfunc='count'))
```

コメント

| カテゴリー | エリア | コメント |
|--------|-------|------|
| A_レジャー | 01_登別 | 1000 |
| | 02_草津 | 1000 |
| | 03_箱根 | 1000 |
| | 04_道後 | 1000 |

データの読み込み

- ① セルをクリックして選択し、再生ボタンを押す
→ ファイルを DataFrame 型に読み込むことができます

集計表の作成

- ② セルをクリックして選択し、再生ボタンを押す
→ 集計表(1)が作成されます

演習:

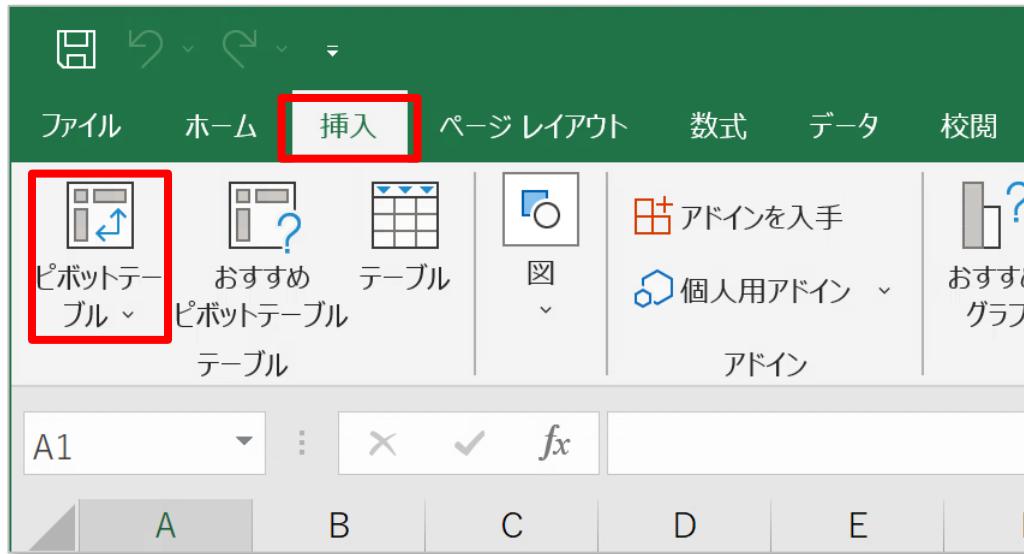
集計表(2)~(9)についても、同様の手順で集計表を作成し、分析対象データの特徴や傾向を読み取り、データを理解する

参考 — EXCEL を使った集計

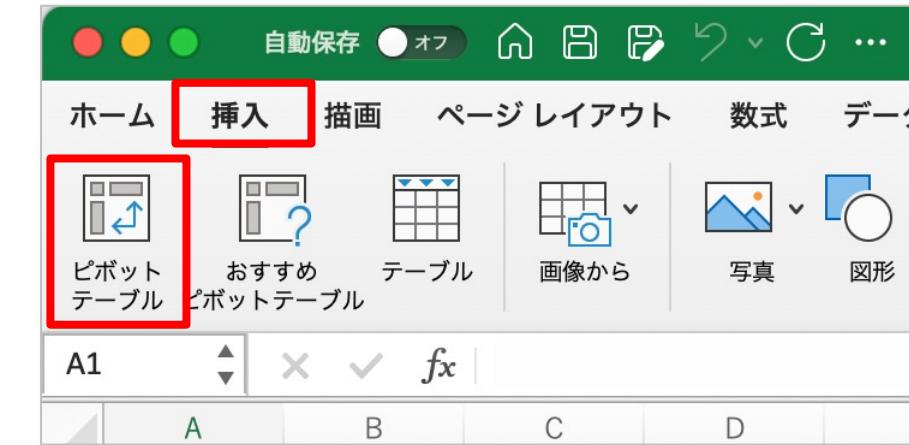
● EXCEL のピボットテーブルを使ってデータを集計する

- ① ファイル **rakuten-1000-2020-2021.xlsx** を開く
- ② A～R 列を選択し、ピボットテーブルを作成する
- ③ [挿入] タブ [テーブル] グループの [ピボットテーブル] ボタンをクリックする

Windows



Mac



データ理解 — 集計例

①件数 (エリア別)

| 行ラベル | 個数 / コメン |
|-----------|--------------|
| ■ A_レジャー | 5000 |
| 01_登別 | 1000 |
| 02_草津 | 1000 |
| 03_箱根 | 1000 |
| 04_道後 | 1000 |
| 05_湯布院 | 1000 |
| ■ B_ビジネス | 5000 |
| 06_札幌 | 1000 |
| 07_名古屋 | 1000 |
| 08_東京 | 1000 |
| 09_大阪 | 1000 |
| 10_福岡 | 1000 |
| 総計 | 10000 |

②投稿者の傾向 (年代別x性別)

| 行ラベル | 個数 / コメン | 列ラベル | 男性 | 女性 | na | 総計 |
|-----------|----------|------|---------------|---------------|---------------|----------------|
| 10代 | | | 0.07% | 0.01% | 0.00% | 0.08% |
| 20代 | | | 0.88% | 1.17% | 0.00% | 2.05% |
| 30代 | | | 2.28% | 2.37% | 0.00% | 4.65% |
| 40代 | | | 5.19% | 3.54% | 0.00% | 8.73% |
| 50代 | | | 7.83% | 4.12% | 0.00% | 11.95% |
| 60代 | | | 4.88% | 2.05% | 0.00% | 6.93% |
| 70代 | | | 1.00% | 0.31% | 0.00% | 1.31% |
| 80代 | | | 0.09% | 0.02% | 0.00% | 0.11% |
| na | | | 0.00% | 0.00% | 64.19% | 64.19% |
| 総計 | | | 22.22% | 13.59% | 64.19% | 100.00% |

③投稿者の傾向 (性別xカテゴリ別)

| 行ラベル | 個数 / コメン | 列ラベル | A_レジャー | B_ビジネス | 総計 |
|-----------|----------|------|----------------|----------------|----------------|
| 男性 | | | 22.26% | 22.18% | 22.22% |
| 女性 | | | 15.22% | 11.96% | 13.59% |
| na | | | 62.52% | 65.86% | 64.19% |
| 総計 | | | 100.00% | 100.00% | 100.00% |

データ理解 — 集計例

④投稿者の傾向 (性別xカテゴリーエリア別)

| 個数 / コメント | 列ラベル | A_レジャー 集計 | | | | | | | | | | B_ビジネス 集計 | | | 総計 |
|-----------|--------|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|-----------|---------|---------|----|
| | | 01_登別 | 02_草津 | 03_箱根 | 04_道後 | 05_湯布院 | 06_札幌 | 07_名古屋 | 08_東京 | 09_大阪 | 10_福岡 | | | | |
| 男性 | A_レジャー | 26.70% | 23.90% | 16.30% | 24.80% | 19.60% | 22.26% | 24.70% | 22.20% | 20.50% | 20.20% | 23.30% | 22.18% | 22.22% | |
| 女性 | A_レジャー | 13.10% | 15.60% | 16.00% | 12.50% | 18.90% | 15.22% | 12.60% | 10.50% | 12.70% | 11.80% | 12.20% | 11.96% | 13.59% | |
| na | A_レジャー | 60.20% | 60.50% | 67.70% | 62.70% | 61.50% | 62.52% | 62.70% | 67.30% | 66.80% | 68.00% | 64.50% | 65.86% | 64.19% | |
| 総計 | | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | |

⑤投稿者の傾向 (年代別xカテゴリーエリア別)

| 個数 / コメント | 列ラベル | A_レジャー 集計 | | | | | | | | | | B_ビジネス 集計 | | | 総計 |
|-----------|--------|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|-----------|---------|---------|----|
| | | 01_登別 | 02_草津 | 03_箱根 | 04_道後 | 05_湯布院 | 06_札幌 | 07_名古屋 | 08_東京 | 09_大阪 | 10_福岡 | | | | |
| 10代 | A_レジャー | 0.00% | 0.00% | 0.30% | 0.00% | 0.00% | 0.06% | 0.20% | 0.00% | 0.20% | 0.10% | 0.00% | 0.10% | 0.08% | |
| 20代 | A_レジャー | 1.60% | 3.70% | 2.80% | 1.40% | 2.70% | 2.44% | 1.30% | 2.00% | 1.40% | 2.10% | 1.50% | 1.66% | 2.05% | |
| 30代 | A_レジャー | 5.10% | 4.60% | 5.60% | 4.40% | 5.80% | 5.10% | 6.40% | 3.50% | 3.40% | 3.80% | 3.90% | 4.20% | 4.65% | |
| 40代 | A_レジャー | 9.80% | 10.00% | 6.20% | 8.00% | 8.60% | 8.52% | 9.20% | 9.20% | 9.80% | 7.60% | 8.90% | 8.94% | 8.73% | |
| 50代 | A_レジャー | 13.20% | 11.40% | 8.90% | 13.10% | 11.80% | 11.68% | 12.40% | 11.10% | 11.60% | 11.30% | 14.70% | 12.22% | 11.95% | |
| 60代 | A_レジャー | 7.70% | 8.30% | 5.70% | 8.80% | 8.60% | 7.82% | 7.00% | 6.20% | 5.70% | 5.80% | 5.50% | 6.04% | 6.93% | |
| 70代 | A_レジャー | 2.30% | 1.10% | 2.80% | 1.50% | 0.90% | 1.72% | 0.70% | 0.50% | 1.00% | 1.30% | 1.00% | 0.90% | 1.31% | |
| 80代 | A_レジャー | 0.10% | 0.40% | 0.00% | 0.10% | 0.10% | 0.14% | 0.10% | 0.20% | 0.10% | 0.00% | 0.00% | 0.08% | 0.11% | |
| na | A_レジャー | 60.20% | 60.50% | 67.70% | 62.70% | 61.50% | 62.52% | 62.70% | 67.30% | 66.80% | 68.00% | 64.50% | 65.86% | 64.19% | |
| 総計 | | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | |

データ理解 — 集計例

⑥投稿者の傾向 (同行者別xカテゴリ-エリア別)

| 個数 / コメント 行ラベル | 列ラベル 01_登別 | A_レジャー 集計 | | | | | | | | | | B_ビジネス 集計 | | | 総計 |
|-------------------|---------------|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|-----------|---------|---------|----|
| | | 02_草津 | 03_箱根 | 04_道後 | 05_湯布院 | 06_札幌 | 07_名古屋 | 08_東京 | 09_大阪 | 10_福岡 | | | | | |
| 一人 | 28.40% | 15.90% | 13.90% | 47.90% | 16.40% | 24.50% | 60.20% | 65.20% | 67.20% | 55.70% | 55.20% | 60.70% | | 42.60% | |
| 家族 | 59.10% | 61.80% | 68.00% | 42.00% | 65.90% | 59.36% | 27.90% | 24.20% | 21.50% | 30.20% | 32.70% | | 27.30% | 43.33% | |
| 恋人 | 5.40% | 13.00% | 10.20% | 4.60% | 10.40% | 8.72% | 5.50% | 4.70% | 4.10% | 5.50% | 4.10% | | 4.78% | 6.75% | |
| 友達 | 4.60% | 8.30% | 6.40% | 3.50% | 6.20% | 5.80% | 3.90% | 3.50% | 4.10% | 6.50% | 5.80% | | 4.76% | 5.28% | |
| 仕事仲間 | 2.00% | 0.60% | 1.00% | 1.40% | 0.40% | 1.08% | 1.80% | 1.70% | 1.90% | 1.60% | 1.80% | | 1.76% | 1.42% | |
| その他 | 0.50% | 0.40% | 0.50% | 0.60% | 0.70% | 0.54% | 0.70% | 0.70% | 1.20% | 0.50% | 0.40% | | 0.70% | 0.62% | |
| 総計 | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | |

⑦投稿者の傾向 (年代別xカテゴリ-エリア別)

| 個数 / コメント 行ラベル | 列ラベル 01_登別 | A_レジャー 集計 | | | | | | | | | | B_ビジネス 集計 | | | 総計 |
|-------------------|---------------|-----------|---------|---------|---------|---------|---------|---------|---------|---------|---------|-----------|---------|---------|----|
| | | 02_草津 | 03_箱根 | 04_道後 | 05_湯布院 | 06_札幌 | 07_名古屋 | 08_東京 | 09_大阪 | 10_福岡 | | | | | |
| 5 | 41.30% | 49.60% | 48.50% | 46.50% | 67.50% | 50.68% | 44.50% | 40.40% | 41.10% | 47.30% | 42.60% | | 43.18% | 46.93% | |
| 4 | 39.20% | 34.20% | 36.00% | 39.40% | 23.70% | 34.50% | 39.50% | 43.60% | 41.20% | 36.90% | 39.80% | | 40.20% | 37.35% | |
| 3 | 11.30% | 9.70% | 8.20% | 9.10% | 5.50% | 8.76% | 9.90% | 10.40% | 11.30% | 10.50% | 12.90% | | 11.00% | 9.88% | |
| 2 | 5.00% | 4.10% | 4.60% | 3.60% | 2.40% | 3.94% | 3.80% | 2.80% | 3.60% | 3.00% | 2.80% | | 3.20% | 3.57% | |
| 1 | 3.20% | 2.40% | 2.70% | 1.40% | 0.90% | 2.12% | 2.30% | 2.80% | 2.80% | 2.30% | 1.90% | | 2.42% | 2.27% | |
| 総計 | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | 100.00% | |

データ理解 — 集計例

⑧-a 数値評価の平均 (エリア別×数値評価別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | 平均 / 食事 | 平均 / 総合 |
|----------|-----------|---------|---------|-------------|---------|---------|---------|
| ■ A_レジャー | 4.22 | 4.28 | 4.11 | 4.01 | 4.29 | 4.26 | 4.28 |
| 01_登別 | 4.03 | 4.27 | 3.95 | 3.88 | 4.31 | 4.08 | 4.10 |
| 02_草津 | 4.19 | 4.28 | 4.03 | 3.92 | 4.31 | 4.15 | 4.25 |
| 03_箱根 | 4.22 | 4.15 | 4.12 | 3.97 | 4.22 | 4.28 | 4.23 |
| 04_道後 | 4.16 | 4.41 | 4.10 | 4.00 | 4.09 | 4.21 | 4.26 |
| 05_湯布院 | 4.52 | 4.28 | 4.36 | 4.27 | 4.50 | 4.57 | 4.55 |
| ■ B_ビジネス | 4.00 | 4.34 | 4.10 | 3.92 | 3.82 | 4.06 | 4.19 |
| 06_札幌 | 3.99 | 4.37 | 4.09 | 3.92 | 3.81 | 4.17 | 4.20 |
| 07_名古屋 | 3.98 | 4.26 | 4.06 | 3.92 | 3.82 | 3.99 | 4.16 |
| 08_東京 | 3.97 | 4.34 | 4.11 | 3.91 | 3.73 | 3.99 | 4.14 |
| 09_大阪 | 4.06 | 4.34 | 4.14 | 3.96 | 3.86 | 4.12 | 4.24 |
| 10_福岡 | 4.01 | 4.40 | 4.11 | 3.89 | 3.85 | 4.02 | 4.18 |

⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | 平均 / 食事 | 平均 / 総合 |
|--------|-----------|---------|---------|-------------|---------|---------|---------|
| A_レジャー | 4.22 | 4.28 | 4.11 | 4.01 | 4.29 | 4.26 | 4.28 |
| B_ビジネス | 4.00 | 4.34 | 4.10 | 3.92 | 3.82 | 4.06 | 4.19 |

データ理解 — 集計例

⑨-a 数値評価の平均 (20~30代, 性別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | 平均 / 食事 | 平均 / 総合 |
|----------|-----------|---------|---------|-------------|---------|---------|---------|
| ■ A_レジャー | 4.39 | 4.34 | 4.26 | 4.17 | 4.39 | 4.37 | 4.39 |
| 男性 | 4.30 | 4.28 | 4.20 | 4.15 | 4.35 | 4.27 | 4.34 |
| 女性 | 4.48 | 4.40 | 4.32 | 4.19 | 4.43 | 4.46 | 4.43 |
| ■ B_ビジネス | 4.16 | 4.32 | 4.09 | 4.01 | 3.95 | 4.20 | 4.19 |
| 男性 | 3.90 | 4.17 | 3.86 | 3.80 | 3.73 | 4.13 | 3.99 |
| 女性 | 4.38 | 4.45 | 4.28 | 4.18 | 4.14 | 4.25 | 4.35 |

⑨-b 数値評価の平均 (40~50代, 性別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | 平均 / 食事 | 平均 / 総合 |
|----------|-----------|---------|---------|-------------|---------|---------|---------|
| ■ A_レジャー | 4.29 | 4.35 | 4.17 | 4.05 | 4.35 | 4.28 | 4.35 |
| 男性 | 4.23 | 4.34 | 4.13 | 3.99 | 4.31 | 4.25 | 4.30 |
| 女性 | 4.36 | 4.37 | 4.23 | 4.15 | 4.41 | 4.33 | 4.43 |
| ■ B_ビジネス | 4.03 | 4.35 | 4.10 | 3.92 | 3.83 | 4.04 | 4.23 |
| 男性 | 3.92 | 4.30 | 4.01 | 3.82 | 3.74 | 3.91 | 4.13 |
| 女性 | 4.25 | 4.46 | 4.29 | 4.14 | 4.01 | 4.28 | 4.44 |

⑨-c 数値評価の平均 (60~90代, 性別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | 平均 / 食事 | 平均 / 総合 |
|----------|-----------|---------|---------|-------------|---------|---------|---------|
| ■ A_レジャー | 4.18 | 4.21 | 4.05 | 3.95 | 4.25 | 4.29 | 4.26 |
| 男性 | 4.11 | 4.21 | 4.00 | 3.92 | 4.25 | 4.26 | 4.24 |
| 女性 | 4.33 | 4.20 | 4.17 | 4.01 | 4.25 | 4.35 | 4.30 |
| ■ B_ビジネス | 3.93 | 4.30 | 4.08 | 3.88 | 3.82 | 3.89 | 4.15 |
| 男性 | 3.90 | 4.28 | 4.07 | 3.85 | 3.73 | 3.85 | 4.15 |
| 女性 | 4.02 | 4.36 | 4.12 | 3.98 | 4.06 | 4.00 | 4.18 |

データ理解 — 集計結果の整理

| 観点 | データの特徴 | テキスト分析時に注意すべき点 |
|----------------|---|--|
| 年代別・性別 | <ul style="list-style-type: none">約60%が年代や性別を表明していない・・ | <ul style="list-style-type: none">レビュー観点がある年代や性別に偏っている可能性・・ |
| 目的別 | <ul style="list-style-type: none">レジャーは家族が多い、ビジネスは一人が多い・・ | <ul style="list-style-type: none">レビューの観点が性別によって偏っている可能性・・ |
| 数値評価 (総合) | <ul style="list-style-type: none">旅行目的によらず評価は高め・・ | <ul style="list-style-type: none">コメントが好評価に偏っている可能性・・ |
| 数値評価 (項目ごと) | <ul style="list-style-type: none">レジャーの評価は、風呂や食事 > 設備や部屋・・ | <ul style="list-style-type: none">旅行目的によって評価の観点や重みが異なっている可能性・・ |
| 全体 | <ul style="list-style-type: none">・・・ | |

- グループワーク (~20:40)
 - データ集計によって発見した、データセットに関する特徴や傾向、テキスト分析時に注意すべき点について、グループ内で討論する
 - 前ページの表を参考に、集計結果から得られた知見を整理する

数値評価で違いを見るのは難しい

【再掲】⑧-a 数値評価の平均 (エリア別×数値評価別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | | | |
|----------|-----------|---------|---------|-------------|---------|------|------|--|
| ■ A_レジャー | 4.22 | 4.28 | 4.11 | 4.01 | 4.29 | 4.26 | 4.28 | |
| 01_登別 | 4.03 | 4.27 | 3.95 | 3.88 | 4.31 | 4.08 | 4.10 | |
| 02_草津 | 4.19 | 4.28 | 4.03 | 3.92 | 4.31 | 4.15 | 4.25 | |
| 03_箱根 | 4.22 | 4.15 | 4.12 | 3.97 | 4.22 | 4.28 | 4.23 | |
| 04_道後 | 4.16 | 4.41 | 4.10 | 4.00 | 4.09 | 4.21 | 4.26 | |
| 05_湯布院 | 4.52 | 4.28 | 4.36 | | | | 4.55 | |
| ■ B_ビジネス | 4.00 | 4.34 | 4.10 | | | | 4.19 | |
| 06_札幌 | 3.99 | 4.37 | 4.09 | | | | 4.20 | |
| 07_名古屋 | 3.98 | 4.26 | 4.06 | 3.92 | 3.82 | | 4.16 | |
| 08_東京 | 3.97 | 4.34 | 4.11 | 3.91 | 3.73 | 3.99 | 4.14 | |
| 09_大阪 | 4.06 | 4.34 | 4.14 | 3.96 | 3.86 | 4.12 | 4.24 | |
| 10_福岡 | 4.01 | 4.40 | | | | 4.02 | 4.18 | |

- ユーザーの8割が4~5の評価、1~2をつけない→本音が見えない

- 同じ点数でもテキストを見れば差異があるかも

- すべての項目に回答する→どこに注目しているかよくわからない

【再掲】⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | 平均 / 食事 | 平均 / 総合 |
|--------|-----------|---------|---------|-------------|---------|---------|---------|
| A_レジャー | 4.22 | 4.28 | 4.11 | 4.01 | 4.29 | 4.26 | 4.28 |
| B_ビジネス | 4.00 | 4.34 | 4.10 | 3.92 | 3.82 | 4.06 | 4.19 |

辻井康一 and 津田和彦「テキストマイニングを用いた宿泊レビューからの注目情報抽出方法」, デジタルプラクティス 3.4 (2012): 289-296.

【再掲】⑧-b 数値評価の平均 (カテゴリ別×数値評価別)

| 行ラベル | 平均 / サービス | 平均 / 立地 | 平均 / 部屋 | 平均 / 設備・アメニ | 平均 / 風呂 | 平均 / 食事 | 平均 / 総合 |
|--------|-----------|---------|---------|-------------|---------|---------|---------|
| A_レジャー | 4.22 | 4.28 | 4.11 | 4.01 | 4.29 | 4.26 | 4.28 |
| B_ビジネス | 4.00 | 4.34 | 4.10 | 3.92 | 3.82 | 4.06 | 4.19 |

● 数値評価のみから違いを見つけるのは難しい！！

- ・ ユーザーの 8割が 4~5 の評価, 1~2をつけない
- ・ ユーザーは 注目の有無に関係なくすべての項目に回答

→ レジャーとビジネスでは、評価すべき項目も異なることを確認した

→ テキストと対応付ければ、同じ点数でも差異があることを確認した

day 2 – レポート課題

- 以下を PDF ファイルで提出してください
 - データ集計により作成した「集計表」のキャプチャ (P.47~51) ※ページ番号は各スライド右下に記載
 - 作成した「集計結果の整理」の表 (P.52) ※ページ番号は各スライド右下に記載
- ※「集計表」のキャプチャは Jupyter の出力でも EXCEL でも構いません
- ※ 何らかの事情で上記2つを提出できない場合、本日の講義の感想を文章で記述してください

| レポート形式 | 提出先 | 期限 |
|--------|--------|----------|
| PDF | manaba | 次回～18:20 |

- 後述する KHCoder を各自の環境にインストールしてください ← このレポート提出は不要

Q&A

次回以降の実習環境について

- 以降の実習では、**KHCoder** (フリーソフトのテキストマイニングツール) を使用します
- KHCoder の利用には Windows OS (10 or 11) が必要になります

| PC の種類 | Windows OSの有無 | 方法 | 備考 |
|---------|------------------------------------|---|--|
| Windows | 有り | Windows PC に KHCoder をインストールして使用する | 最もオススメ |
| | | 全学計算機システムの Windows [※] に KHCoder をインストールして使用する | ※ 利用手順: https://www.u.tsukuba.ac.jp/remote/#vm2win |
| Mac | 仮想環境 [※] 上で動く Windows がある | 仮想環境上の Windows に KHCoder をインストールして使用する | ※ Vmware Fusion や Parallels Desktop などを想定 |
| | なし | 全学計算機システムの Windows [※] に KHCoder をインストールして使用する | ※ 利用手順: https://www.u.tsukuba.ac.jp/remote/#mac2win |
| | | SageMaker Studio Lab 上の Python スクリプト [※] を利用する | ※ 救済的な措置で、一部の KHCoder 機能は未対応です |

KH Coder のインストール (day 3 以降で使用)

- 前ページを参考に、各自で選択した環境(個人PC or 全学計算機システム)に **KHCoder** をインストールしておいてください
- ダウンロードとインストール <https://khcoder.net/dl3.html>



1. **ここをクリックするとダウンロードが始まります**
2. ダウンロードしたファイルを実行 (ダブルクリックし、開いた画面上の「Unzip」ボタンをクリックします)
3. 任意の保存先を指定します (**全学計算機ではCドライブへの保存は禁止されています**)
例: 「Z:¥Desktop¥khcoder3」 (全学の場合)
4. 指定した保存先フォルダにすべてのファイルが解凍されます。解凍された「**kh_coder.exe**」を実行すると KH Coder が起動します。