

スケジュール

day 2

- 講義 – 自然言語処理の最新動向
- 講義 – テキストマイニングの手順
- 講義&演習 – データ理解

day 3

- 講義&演習 – 演習環境の準備
- 講義&演習 – テキスト解析 (1)
- 講義&演習 – テキスト解析 (2)

day 4

- 講義&演習 – テキスト分析 (1)
- 演習 – テキスト分析 (実践編)

day 5

- 演習 – テキスト分析 (実践編)

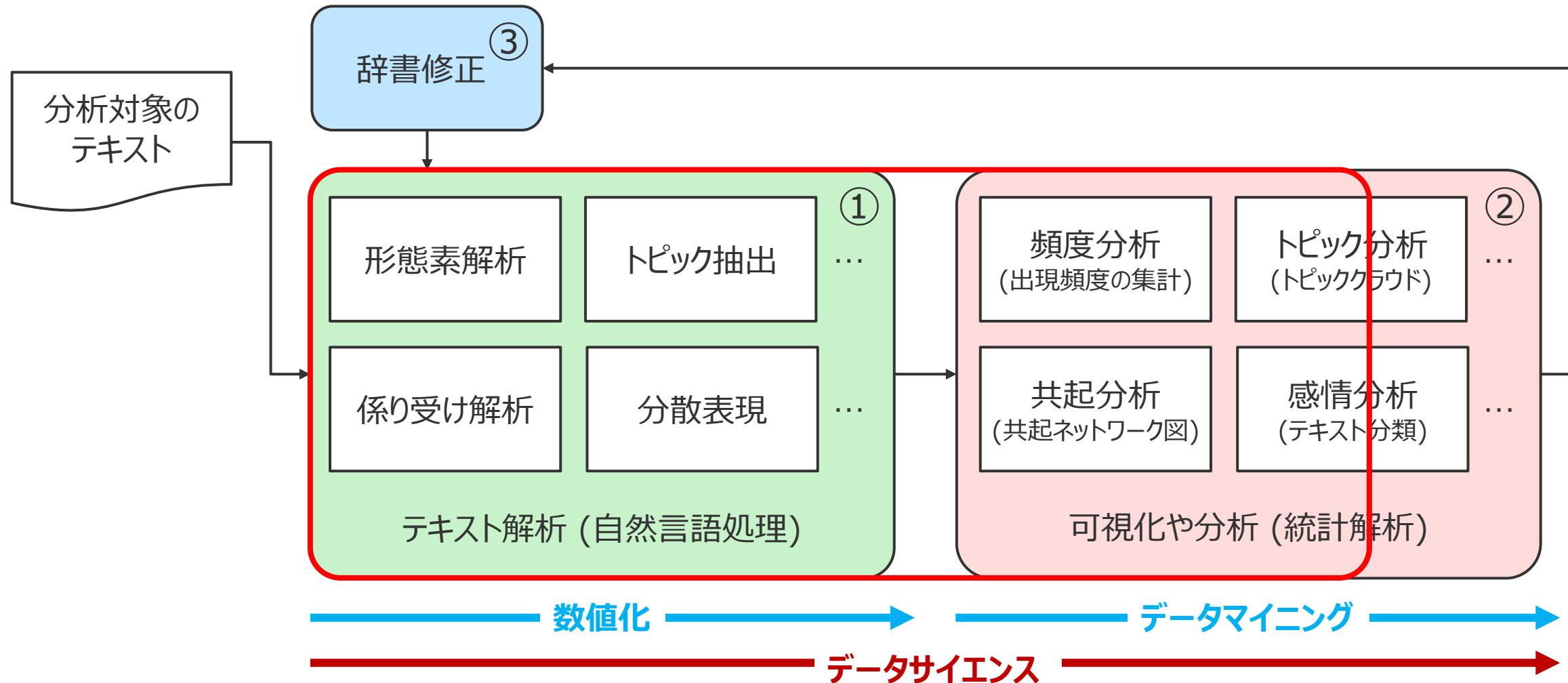
テキスト分析 (1)

(再掲) 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

(再掲) 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)



- 社会調査データを分析する目的で開発されたフリー(~~商用可能~~)のツール

- 高機能かつ~~商用可能~~でフリー
- Rを用いた多変量解析と可視化
- 実装されている分析手法
 - ・ 階層的クラスター分析
 - ・ 多次元尺度構成法(MDS)
 - ・ 対応分析
 - ・ 共起ネットワーク
 - ・ 自己組織化マップ
 - ・ 文書のクラスター分析
 - ・ トピックモデル (LDA)

論文検索サービスも提供 → <http://khcoder.net/bib.html>

研究事例リスト

KH Coderを用いたご研究の成果を発表された際には、書誌情報をフォームにご記入いただけますと幸いです。

出版年 :

著者名 :

キーワード :

ヒット件数 : 0200 / 6135

KH Coderを用いた研究事例のリスト 6135件

※2023/6/16 現在

→1646→2042→2695→3741件→4554件→昨年5355件→6135件)

(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

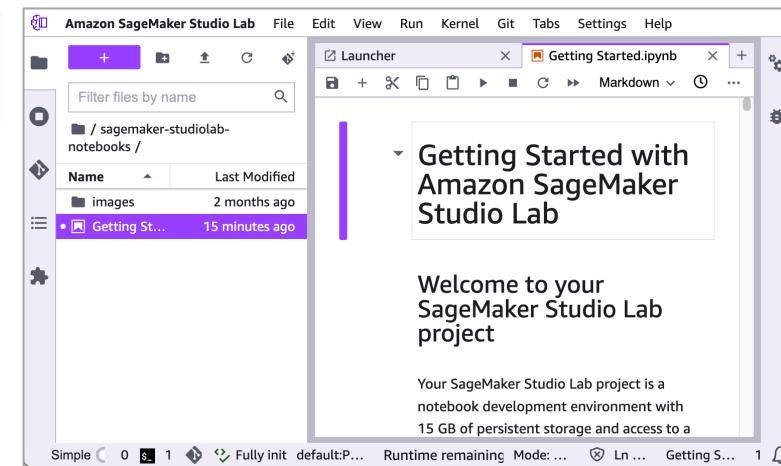
 Colaboratory

<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)

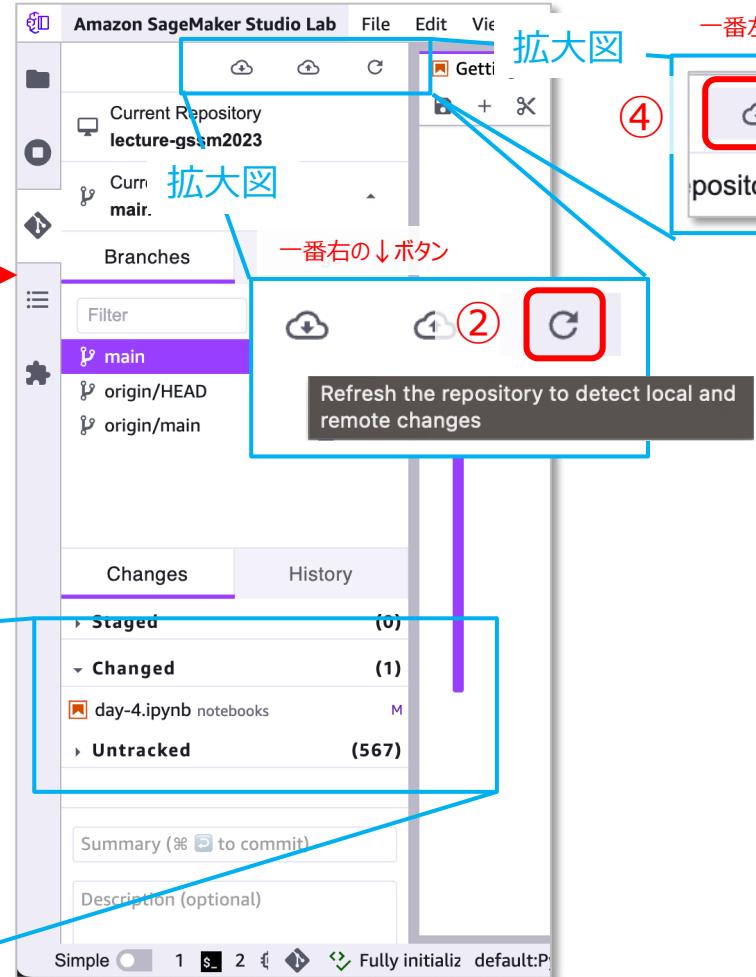
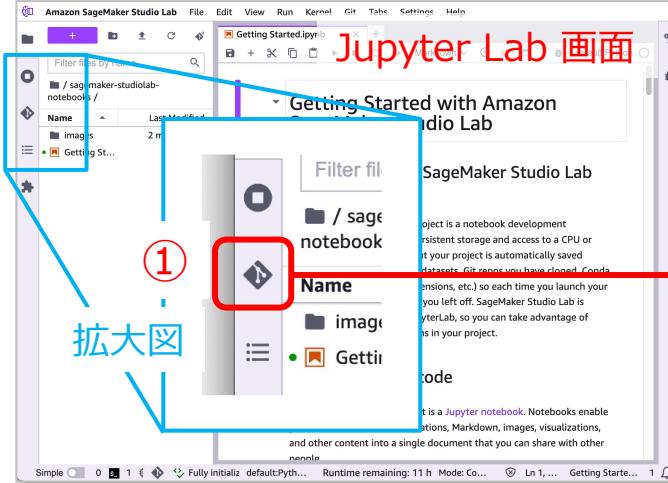
使用するテキスト分析手法

● 主に以下の5種類の分析や可視化手法を利用します

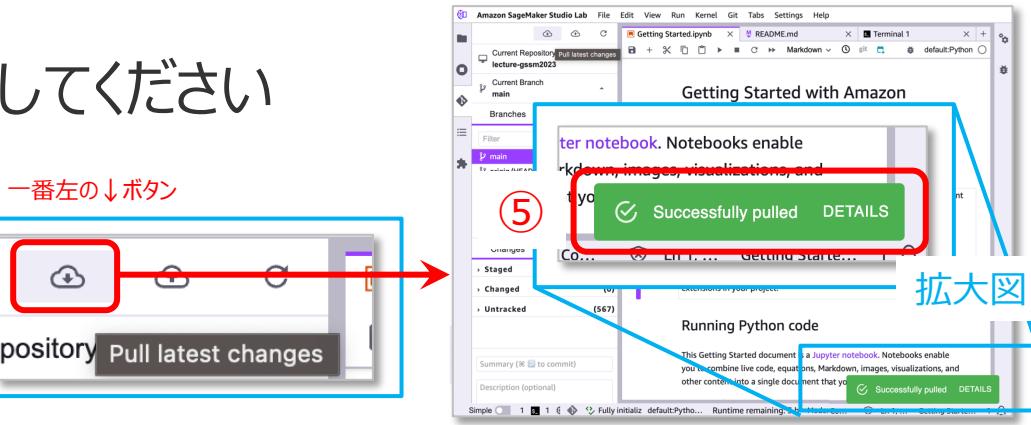
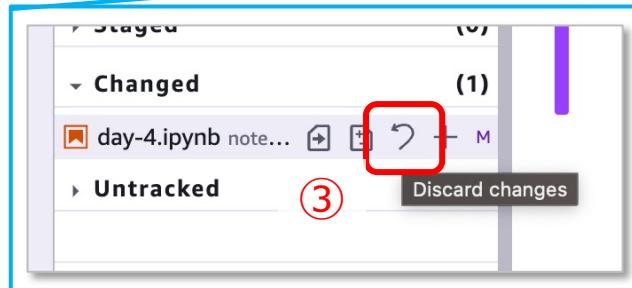
分析・可視化手法	特徴	用途
ワードクラウド	テキスト内で頻出する単語を視覚的に表示する手法。単語の頻度に応じて文字の大きさや色が変わる。	テキスト全体の傾向やキーワードを直感的に把握するために使用する。プレゼンテーション資料などの視覚的な効果も高い。
共起ネットワーク	単語の共起関係(同時に出現する関係)をネットワーク図として表現する手法。ノードが単語をエッジが共起関係を示す。	単語同士の関係性やパターンを分析するために使用する。テキスト内のトピックやテーマの繋がりを理解するために役立つ。
係り受けネットワーク	文中の単語の係り受け関係を視覚化する手法。名詞と形容詞の修飾関係をネットワークで表示する。	文の論理構造や詳細な意味関係を考慮して分析したい場合に使用する。
対応分析プロット	質的データ間の関係を可視化する手法。行列形式のデータを低次元に縮約してプロットする。	外部変数と単語の関連性や相関を調べる際に使用する。
トピックモデル	大量の文書から潜在的なトピックを自動的に抽出する手法。LDA (潜在的ディレクリ分布) アルゴリズムを使用。	大量のテキストデータから主要なトピックを特定するために使用する。ワードクラウドでプロットすることで視覚的な効果も高い。

演習 — テキスト分析 (1)

● Jupyter Lab を起動して、教材を最新化(pull)してください



(例) 競合がある場合のみ 拡大図



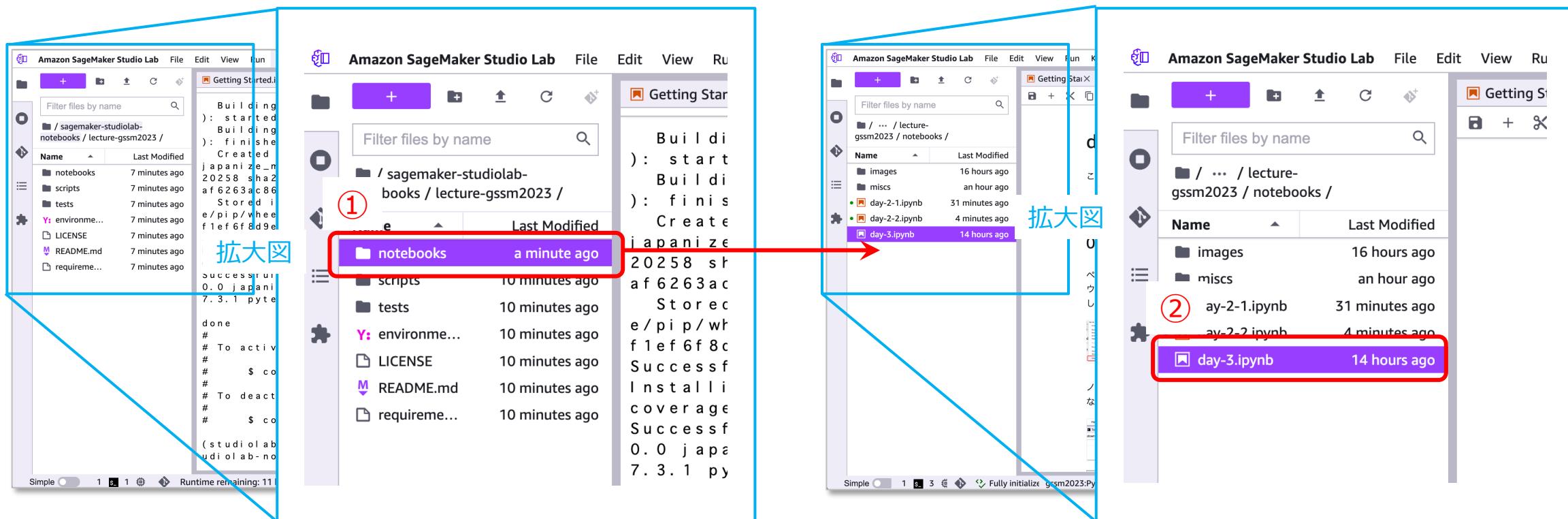
● 教材を最新化する

- ① 「Git」 ボタンを押す
- ② 「Refresh」 ボタンを押す
- ③ もし、競合がある場合(Changedが0でない場合)、
対象のファイルを手動で退避した後、
「Discard changes」 ボタンを押し
て変更を破棄する
- ④ 「Pull latest changes」 を押す
- ⑤ 画面の右下に「Successfully published」が表示されること確認する

演習 — テキスト分析 (1)

● day-4-1.ipynb を開いてください

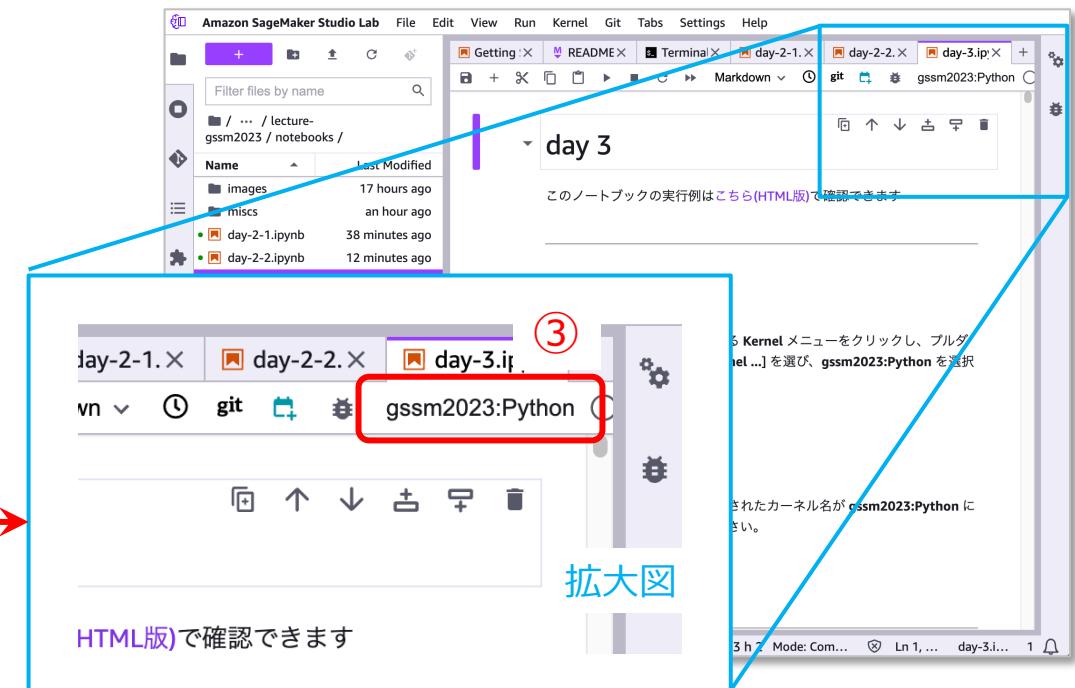
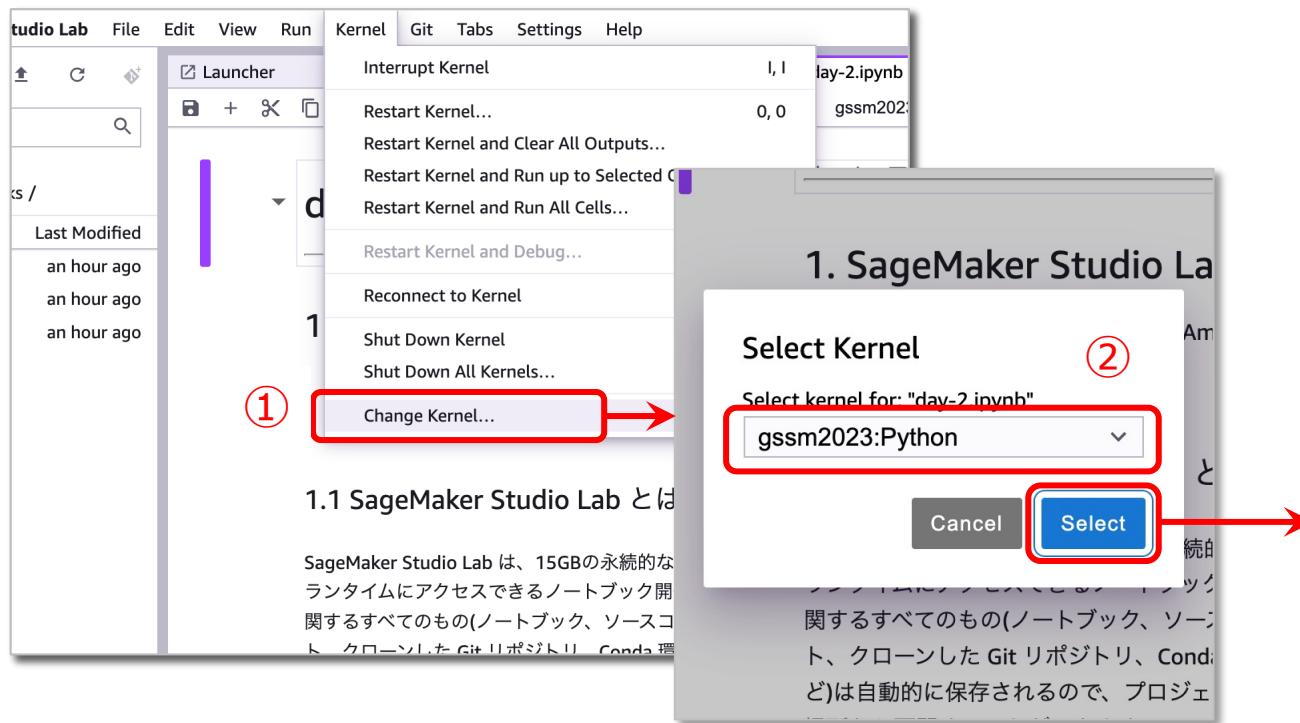
- ① 画面左の **File Browser** から ① **notebooks** をフォルダを開く (既に開いている場合はスキップ)
- ② 次に **day-4-1.ipynb** ノートブックを開く



演習 — テキスト分析 (1)

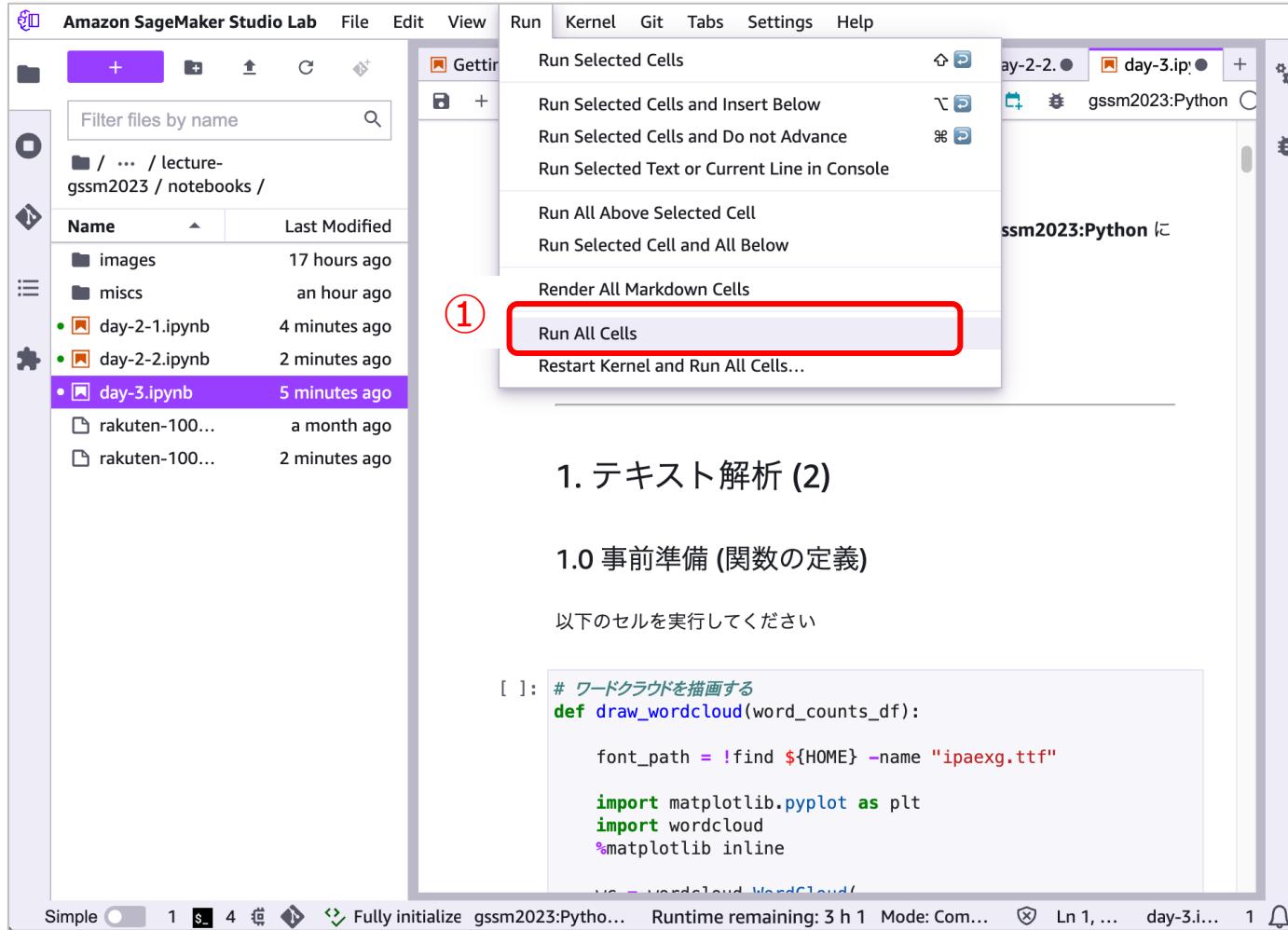
● カーネル gssm2024:Python を選択してください !重要!

- 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「gssm2024:Python」を選択し、「Select」を押す
- ③ 右上隅にカーネル名「gssm2024:Python」が表示されていることを確認する



演習 — テキスト分析 (1)

● テキスト解析と可視化 (day-4-1.ipynb)



The screenshot shows the Amazon SageMaker Studio Lab interface. On the left, there's a file browser with a list of notebooks. In the center, a notebook titled 'day-3.ipynb' is open, showing code for generating a word cloud. A context menu is open over the code cell, with the 'Run All Cells' option highlighted by a red box and a circled number 1.

```
[ ]: # ワードクラウドを描画する
def draw_wordcloud(word_counts_df):
    font_path = !find ${HOME} -name "ipaexg.ttf"
    import matplotlib.pyplot as plt
    import wordcloud
    %matplotlib inline
```

演習:

- ① ページ上部の Run メニューから「Run All Cells」を選ぶ

この後、Step-by-step で解説します

テキスト分析 (実践編)

実践的な分析

- 実践1: カテゴリーやエリアごとのユーザーの注目ポイントを押さえる
- 実践2: カテゴリーやエリアごとのユーザーの注目ポイントの評価の違いを見つける
- 実践3: 高評価のエリアに倣って、低評価のエリアを改善するプランを提案する
→ 注意: プロットによる可視化と宿泊客の生の声(原文)を使って解釈する

例) 実践3のまとめ方

対象エリア	エリアX の評価ポイント	エリアY の課題	エリアYの改善案
エリアX: XXX エリアY: XXX	• 風呂が広い 根拠原文: ... • ...	• エアコンが臭い 根拠原文: ... • ...	• ... • ...