

人文社会ビジネス科学学術院 ビジネス科学研究群 2024年度 春C

# テキストマイニングの実践

## day 4

## 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)

## (前回) day 3 – レポート課題

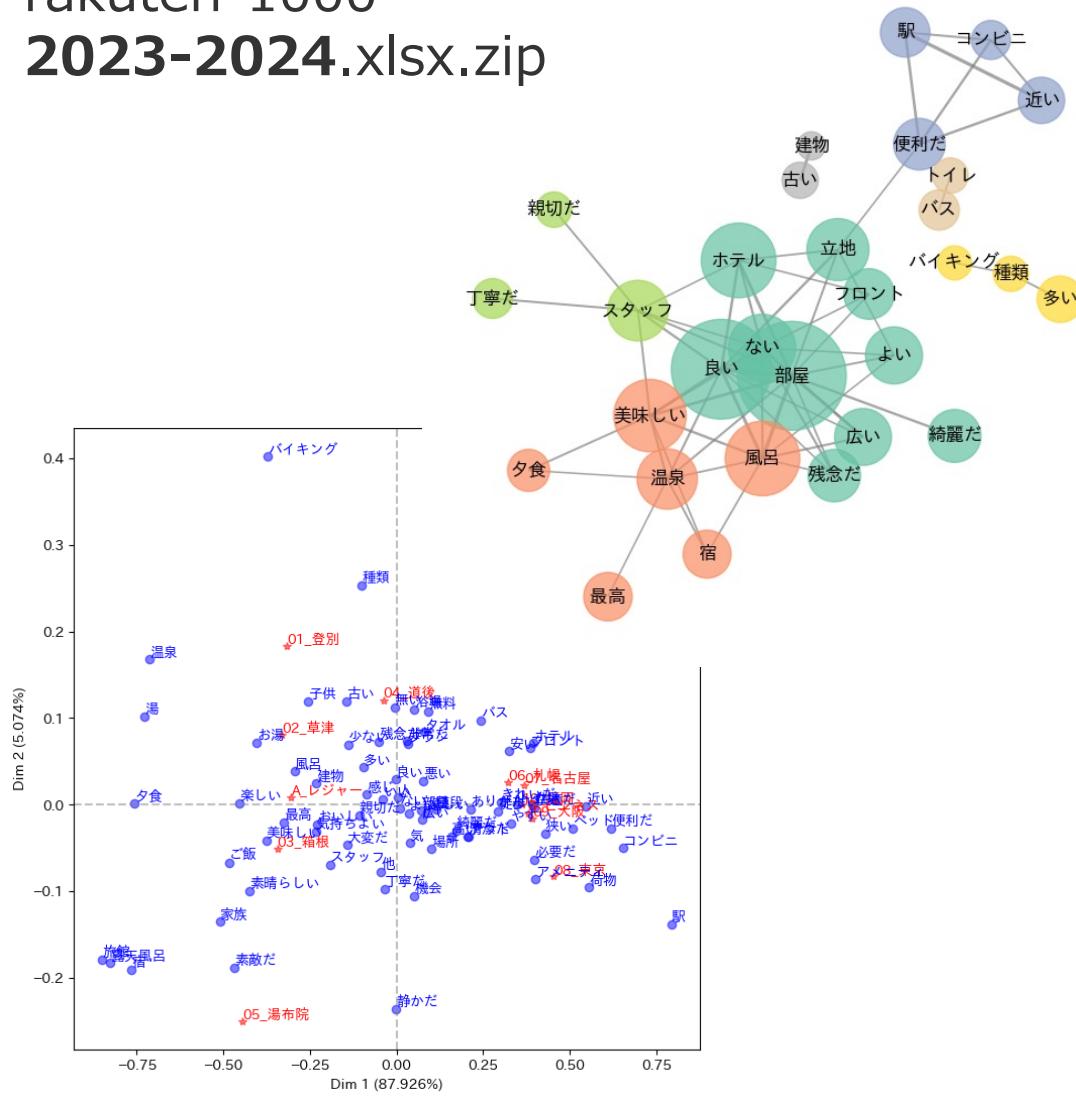
- 以下を PDF ファイルで提出してください
  - ノートブック **day-3-3.ipynb** の末尾にある「【演習】2021~2022 データセット」に従って、別のデータセット (**rakuten-1000-2021-2022.xlsx.zip**) で作図した「共起ネットワーク図」と「対応分析プロット」のキャプチャ

※ 何らかの事情で上記のキャプチャ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

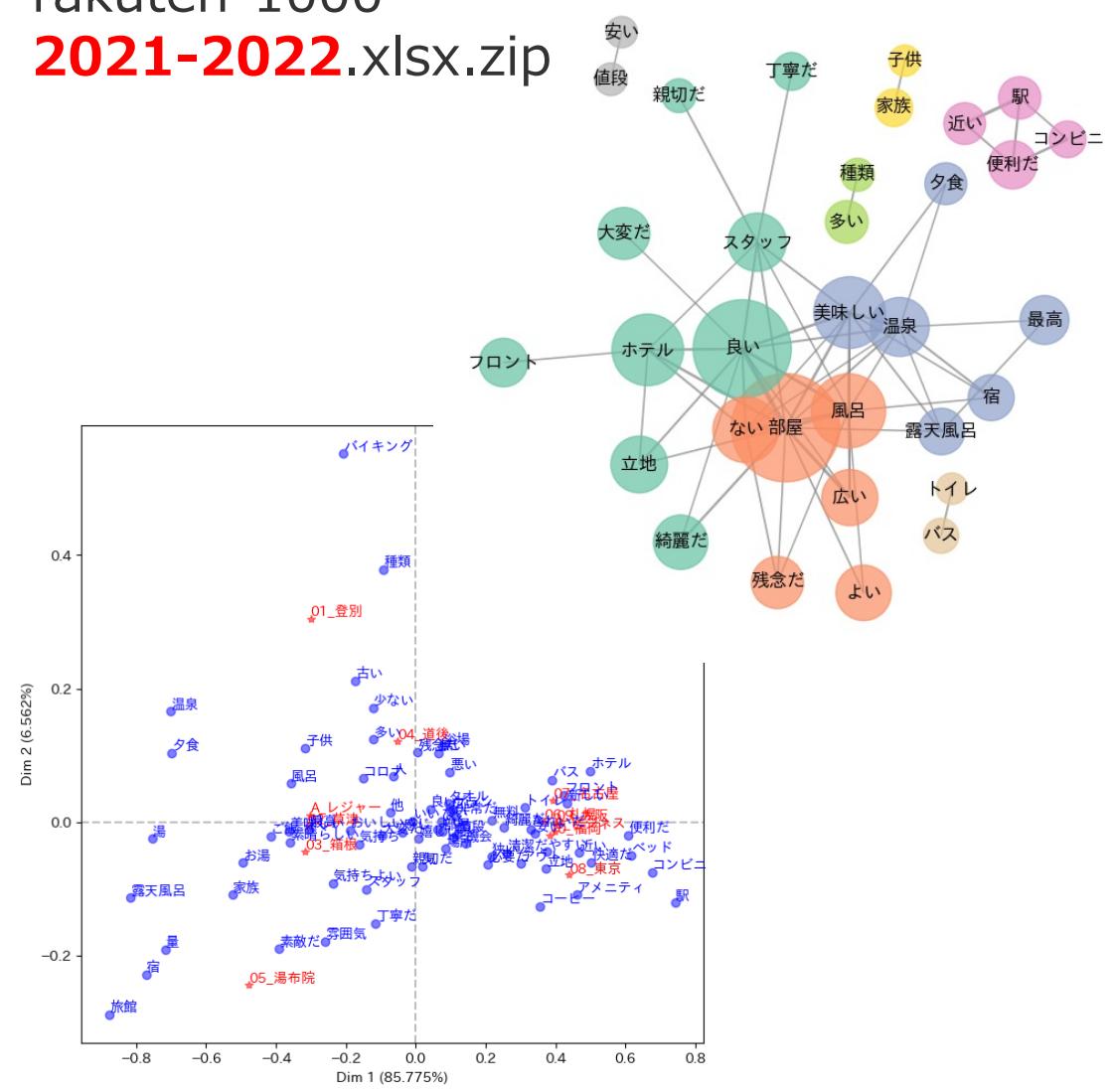
レポート形式	提出先	期限
PDF	manaba	次回～18:20

# (参考) day3 レポート課題のプロット

# rakuten-1000- **2023-2024.xlsx.zip**



# rakuten-1000- **2021-2022.xlsx.zip**



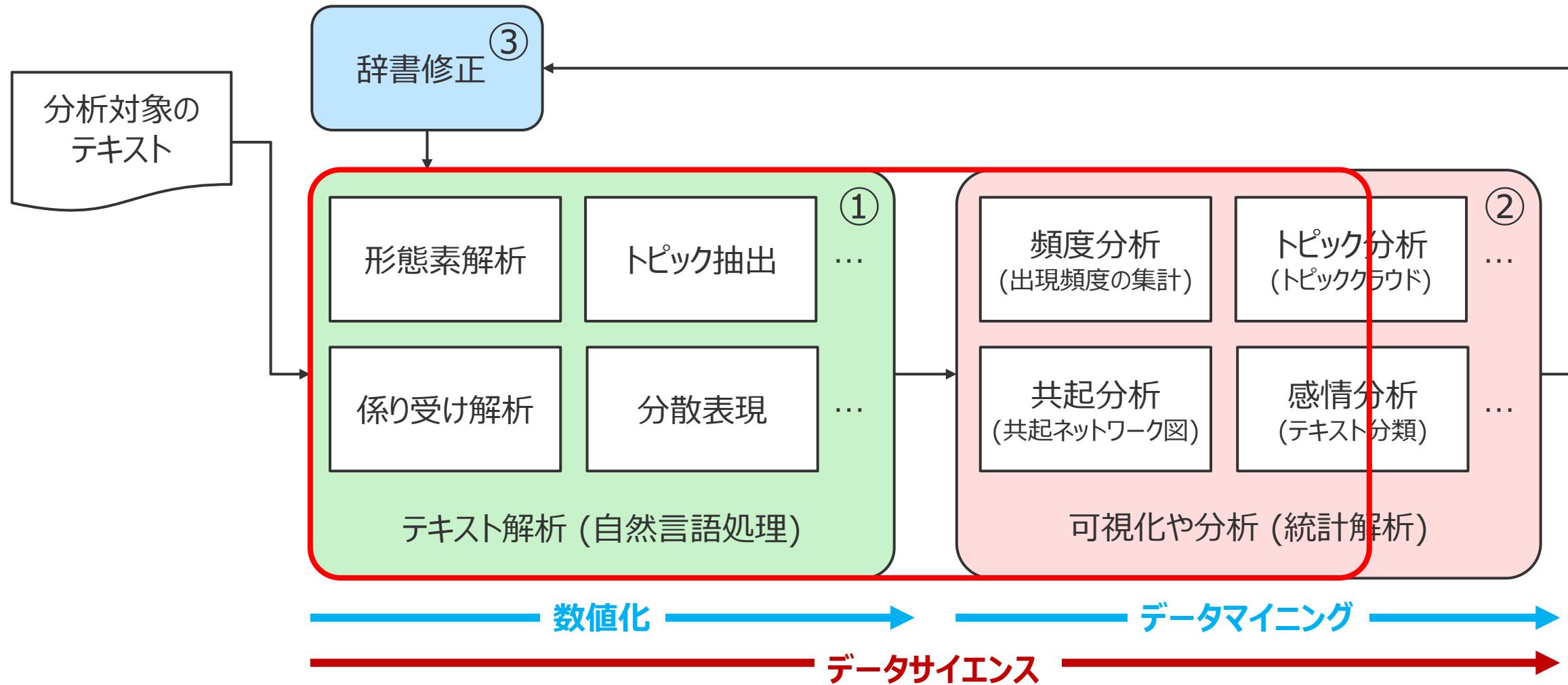
# テキスト分析 (1)

## (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

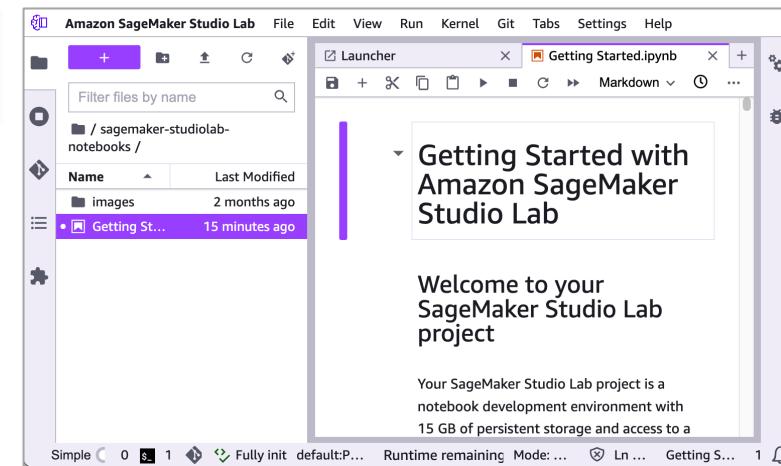
 Colaboratory

<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。

## 使用するテキスト分析手法 (1/5)

## ● ワードクラウド

**特徴:** 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。

**用途:** テキスト  
プレビューで  
共通する注目ポイントが  
分かる

**分析例：宿泊券** まず目<sup>。</sup>ポイントを見つける、○○○○○

# 注目ポイントがカテゴリーごとに異なることが分かる



# 全データのワードクラウド



「A レジャー」のワードクラウド



## ~~「B ビジネス」のワードクラウド~~

## 使用するテキスト分析手法 (2/5)

### ● 共起ネットワーク

**特徴:** 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  
ノードが単語をエッジが共起関係を示す。

**用途:** 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。

**分析例①:** 宿泊者の注目ポイントに対する  
評価を調べる、○○ごとに比較する

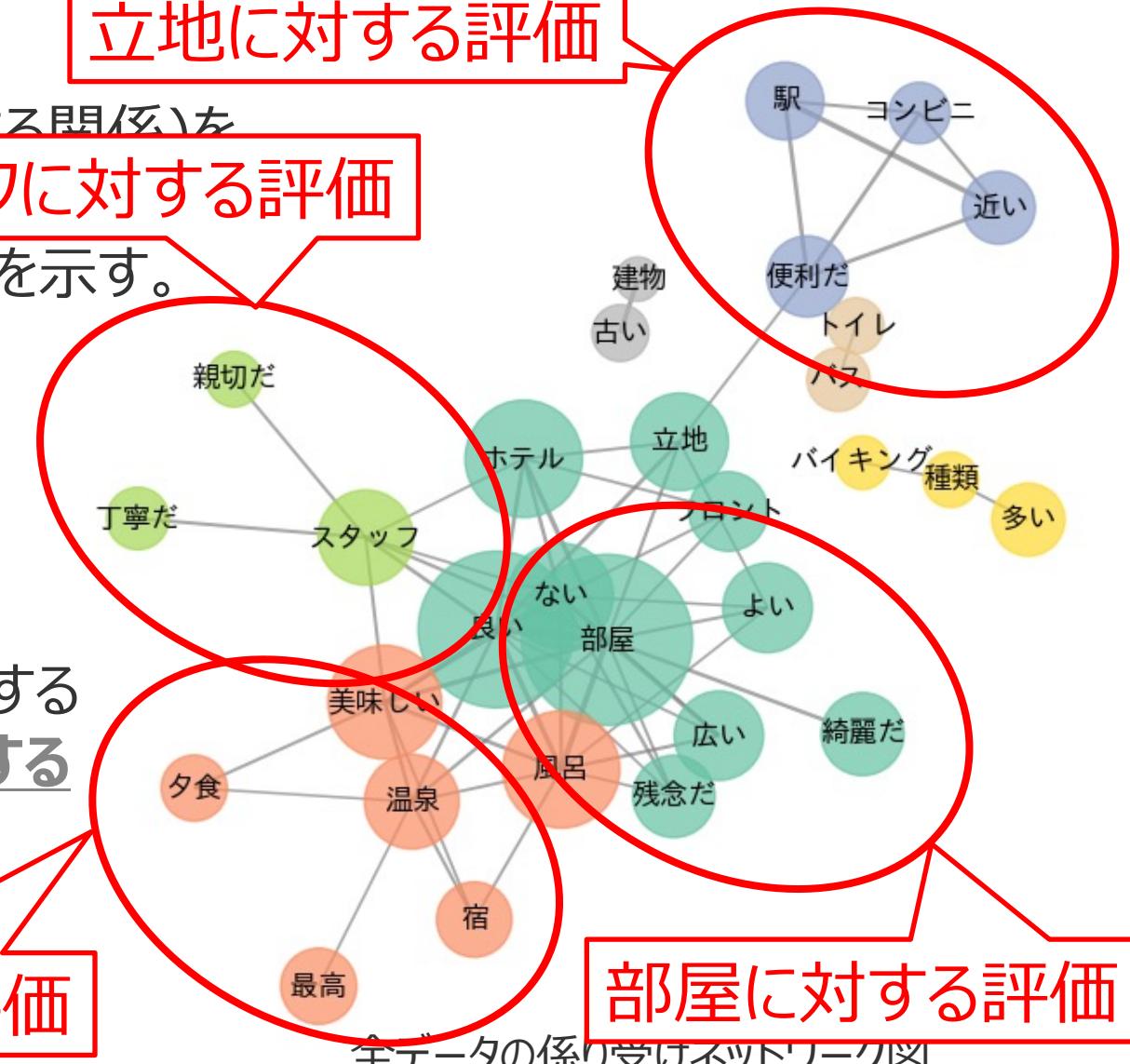
立地に対する評価

スタッフに対する評価

温泉に対する評価

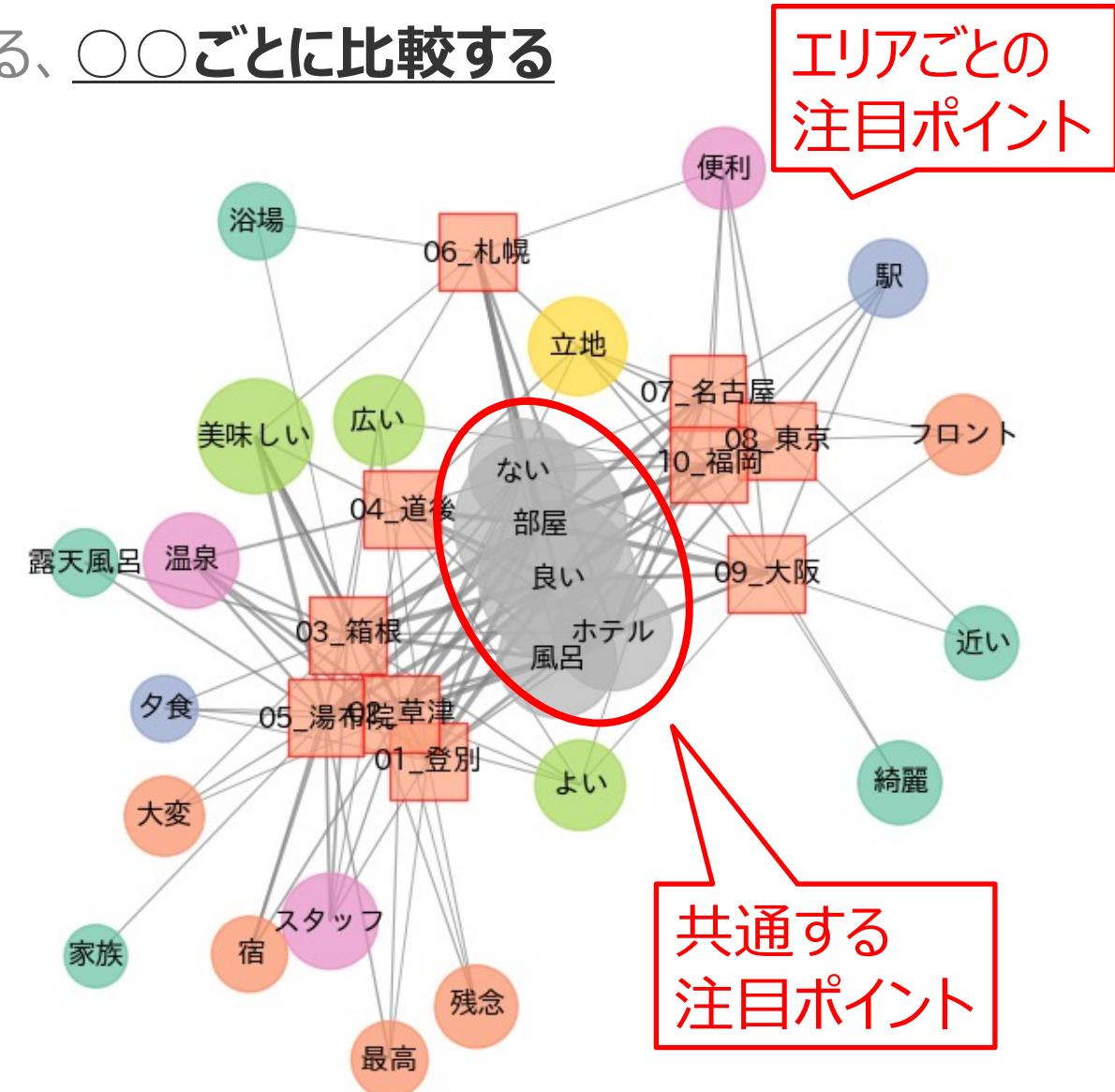
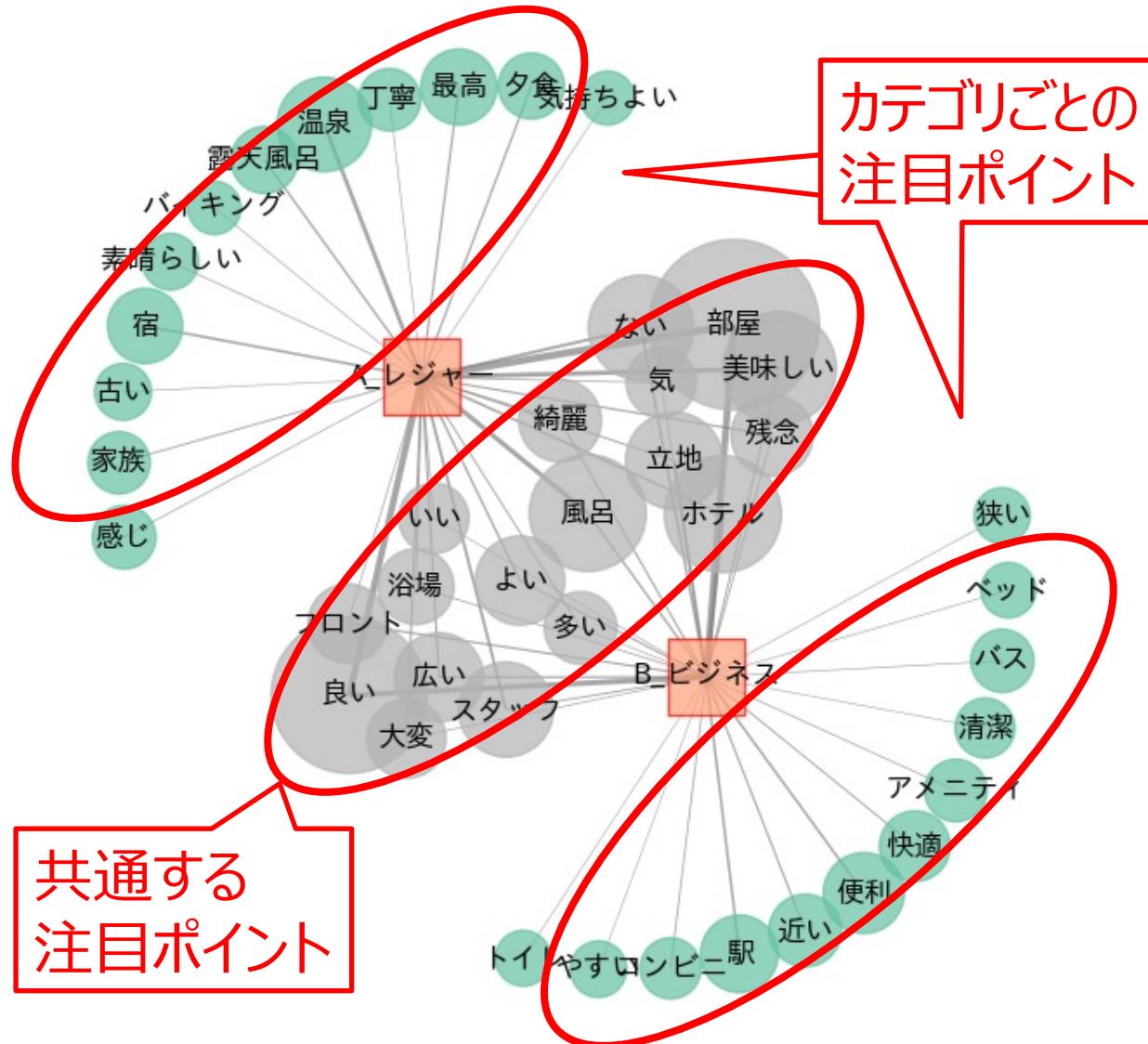
部屋に対する評価

全データの係り受けネットワーク図



# 使用するテキスト分析手法 (2/5)

分析例②：宿泊者の注目ポイントを見つける、〇〇ごとに比較する



## 使用するテキスト分析手法 (3/5)

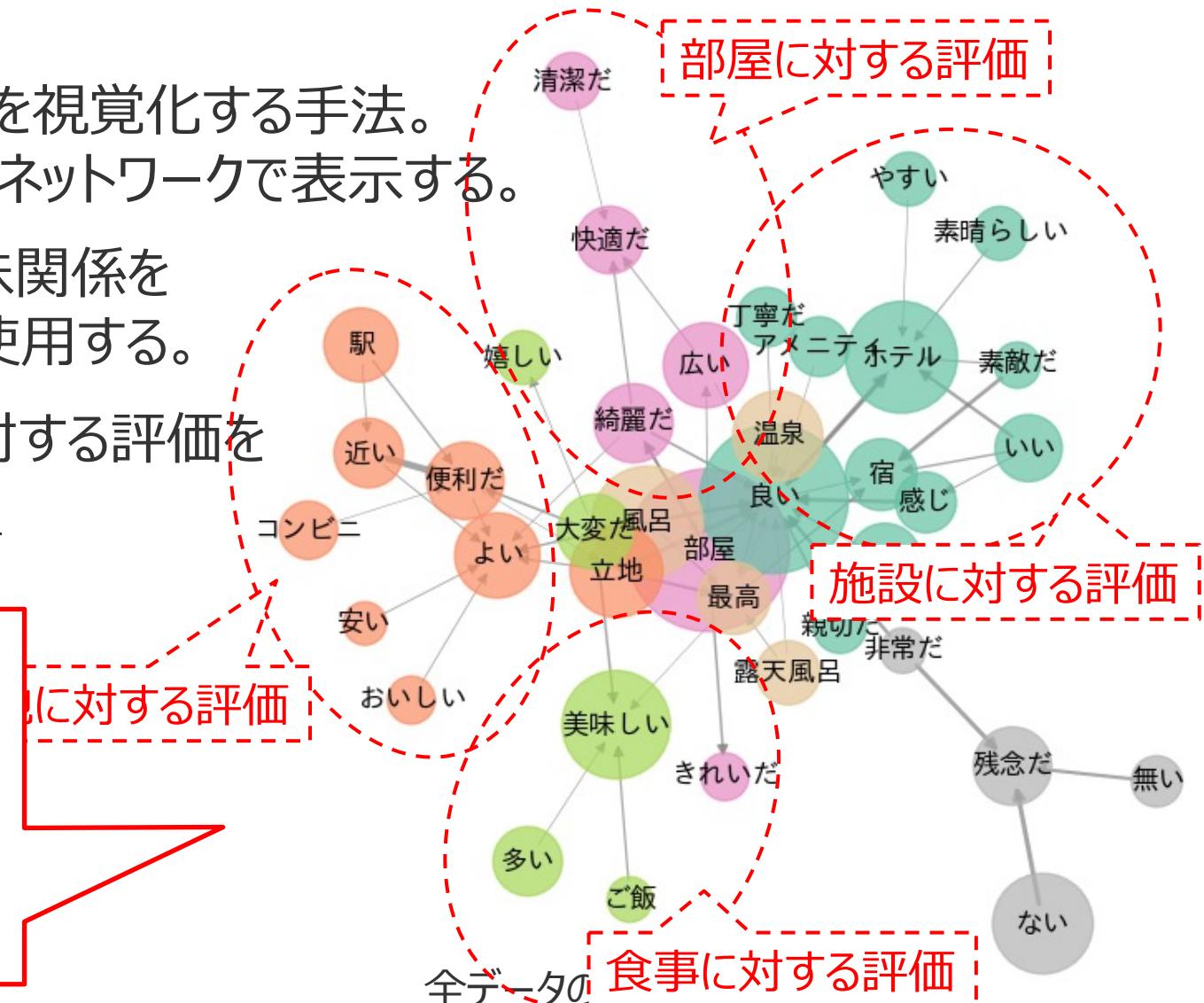
### ● 係り受けネットワーク

**特徴:** 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。

**用途:** 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。

**分析例:** 宿泊者の注目ポイントに対する評価を  
調べる、○○ごとに比較する  
(「共起ネットワーク」と同じ)

- 有向グラフのため、関係の向きまで  
確認できる
- 正確な関連性が表現できる一方で  
共起頻度が減少し、共起パターンが  
現れにくい



# 使用するテキスト分析手法 (4/5)

## ● 対応分析プロット

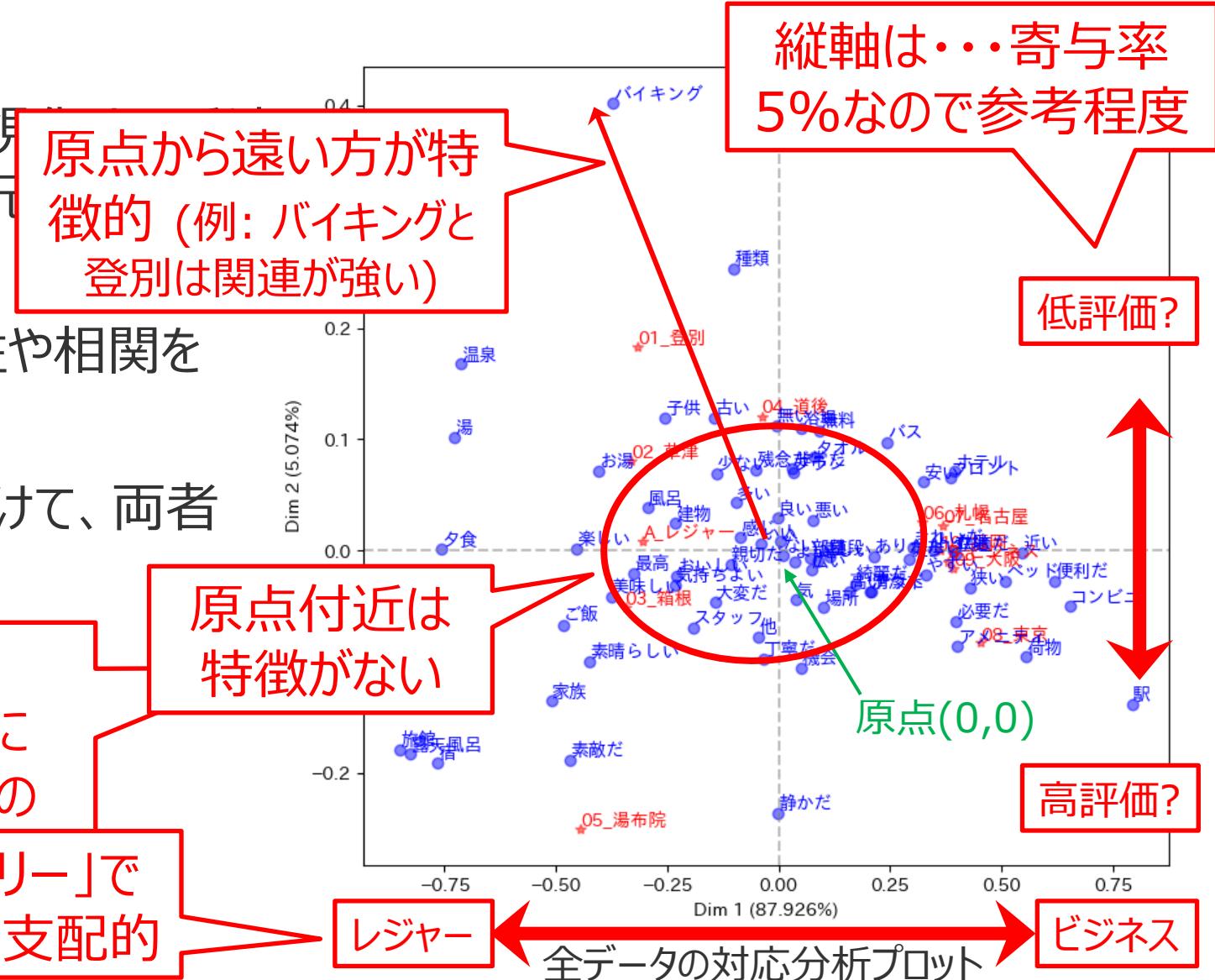
**特徴:** 質的データ間の関係を可視化  
行列形式のデータを低次元  
プロットする。

**用途:** 外部変数と単語の関連性や相関を調べる際に使用する。

**分析例：対照的な2エリアを見つけて、両者の  
の違いを比較する**

第2固有値までの累積寄与率は  
 $87.93 + 5.07 = 93.0\%$  で非常に  
高く、第1,2固有値に対応する軸の  
主成分をすれば

横軸は「カテゴリー」で  
寄与率88%で支配的



# 使用するテキスト分析手法 (5/5)

## ● トピックモデル

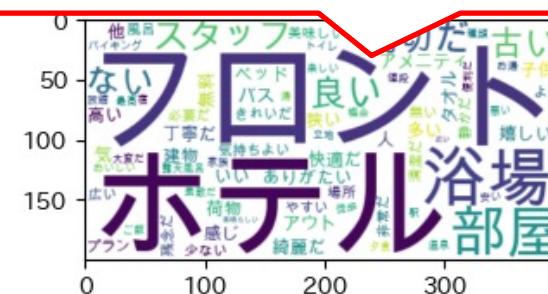
**特徴:** 大量の文書から潜在的なトピックを自動的に抽出する手法。LDA (潜在的ディレクリ分布) アルゴリズムを使用。※線形判別分析のLDAとは全く別もの

**用途:** 大量のテキストデータから主要なトピックを抽出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が高い

**分析例:** 宿泊者の注目トピック(=注目単語の集まり)を自動抽出する、○○ごとでトピックの出現割合を比較する

各トピックが含まれる割合も算出できる

施設のスタッフと設備



施設の立地と利便性



温泉と食事の質



Topic # 4:



Topic # 5:



Topic # 6:



# 使用するテキスト分析手法 (5/5)

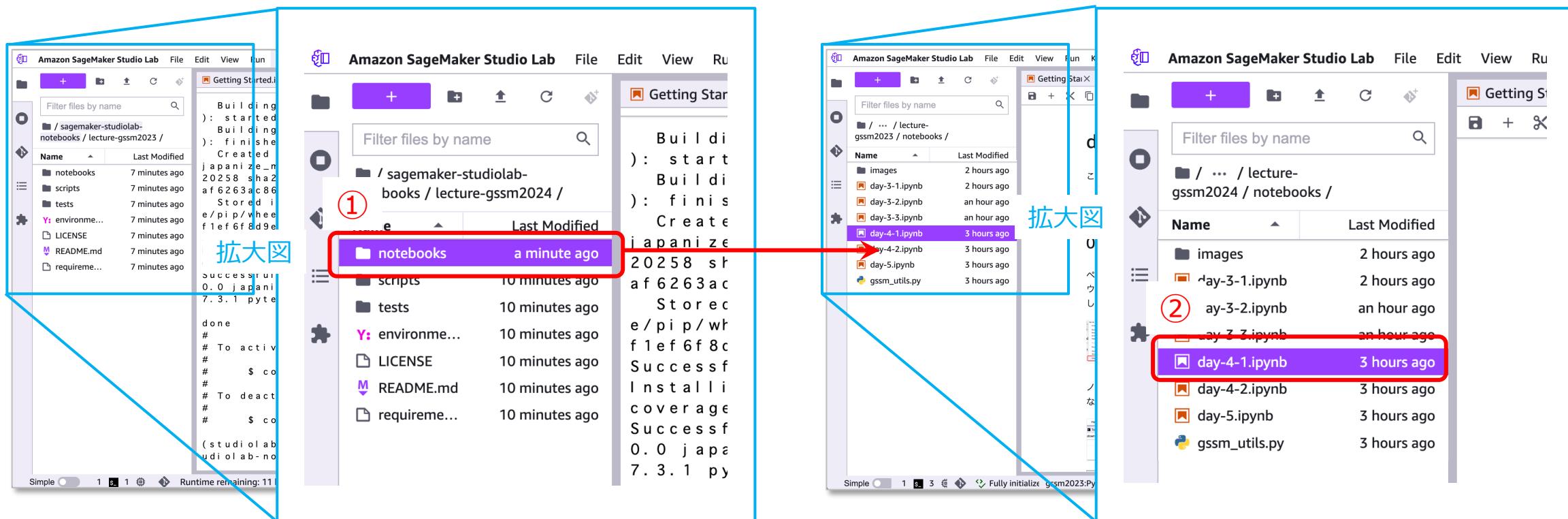
## ● ChatGPT を使ってトピックを説明する

Topic	トピックワード (出現確率 Top20)	トピックの説明 (トピックワードからChatGPTで生成)
# 1	フロント ホテル 浴場 部屋 親切だ 良い スタッフ ない 古い ありがたい バス タオル 綺麗だ アメニティ 荷物 嬉しい 建物 丁寧 だ アウト 多い	<b>施設のスタッフと設備</b> スタッフが親切で、部屋や浴場が清潔だが、建物が古い。
# 2	便利だ 駅 近い コンビニ 立地 良い 部屋 ホテル 徒歩 バス 狹い 必要だ よい ない やすい アメニティ 場所 気 多い 荷物	<b>施設の立地と利便性</b> 駅やコンビニが近く便利だが、部屋が狭いこともある。
# 3	美味しい 温泉 良い 風呂 部屋 最高 宿 露天風呂 家族 夕食 バイキング ご飯 よい 多い 湯 楽しい 子供 お湯 浴場 ない	<b>温泉と食事の質</b> 温泉が良く、食事が美味しい。家族で楽しめる。
# 4	大変だ スタッフ 良い 部屋 素晴らしい 残念だ 美味しい 丁寧 だ ない ホテル 種類 温泉 風呂 無い 素敵だ 宿 夕食 多い 非常だ 少ない	<b>スタッフの対応と設備の質</b> スタッフが丁寧で部屋が清潔だが、設備が不足していることもある。
# 5	良い 部屋 立地 綺麗だ 風呂 ホテル ない トイレ 残念だ 値段 広い ベッド 悪い いい 安い 気 人 高い 感じ 他	<b>部屋の質と価格</b> 部屋や風呂が広く清潔だが、価格が高いと感じる人もいる。
# 6	部屋 広い よい 快適だ 清潔だ きれいだ やすい 機会 おいしい 静かだ 風呂 ホテル 気持ちはいい 大変だ 場所 美味しい 気 アメニティ ない 多い	<b>部屋の広さと快適さ</b> 部屋が広く快適で清潔。静かな環境でリラックスできる。

# 演習 — テキスト分析 (1)

## ● day-4-1.ipynb を開いてください

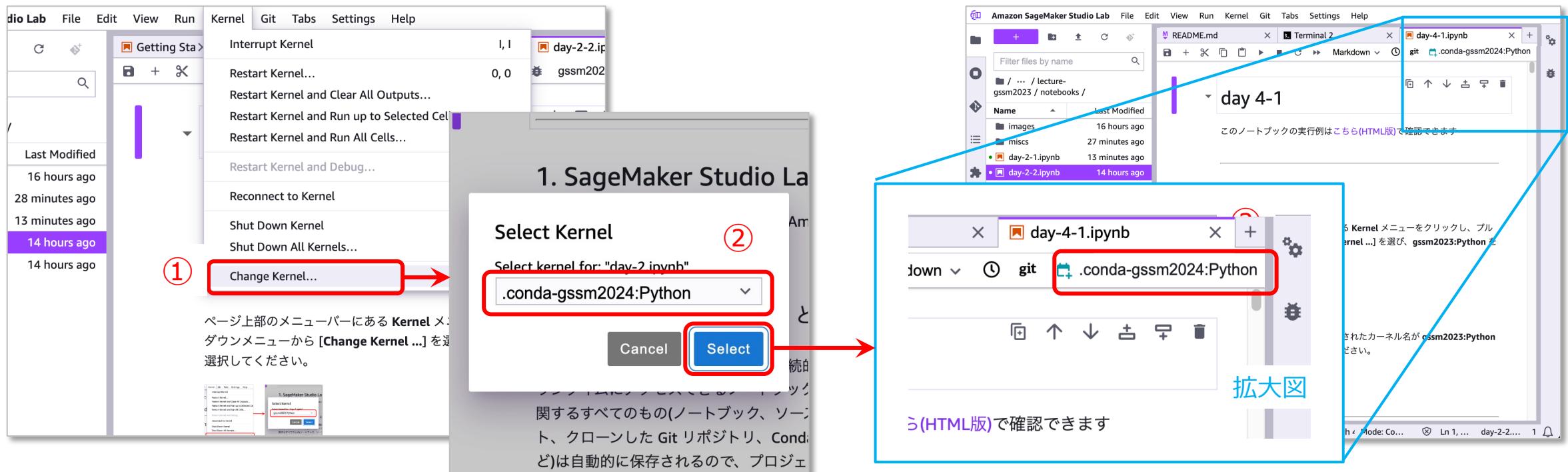
- ① 画面左の **File Browser** から ① **notebooks** をフォルダを開く (既に開いている場合はスキップ)
- ② 次に **day-4-1.ipynb** ノートブックを開く



# 演習 — テキスト分析 (1)

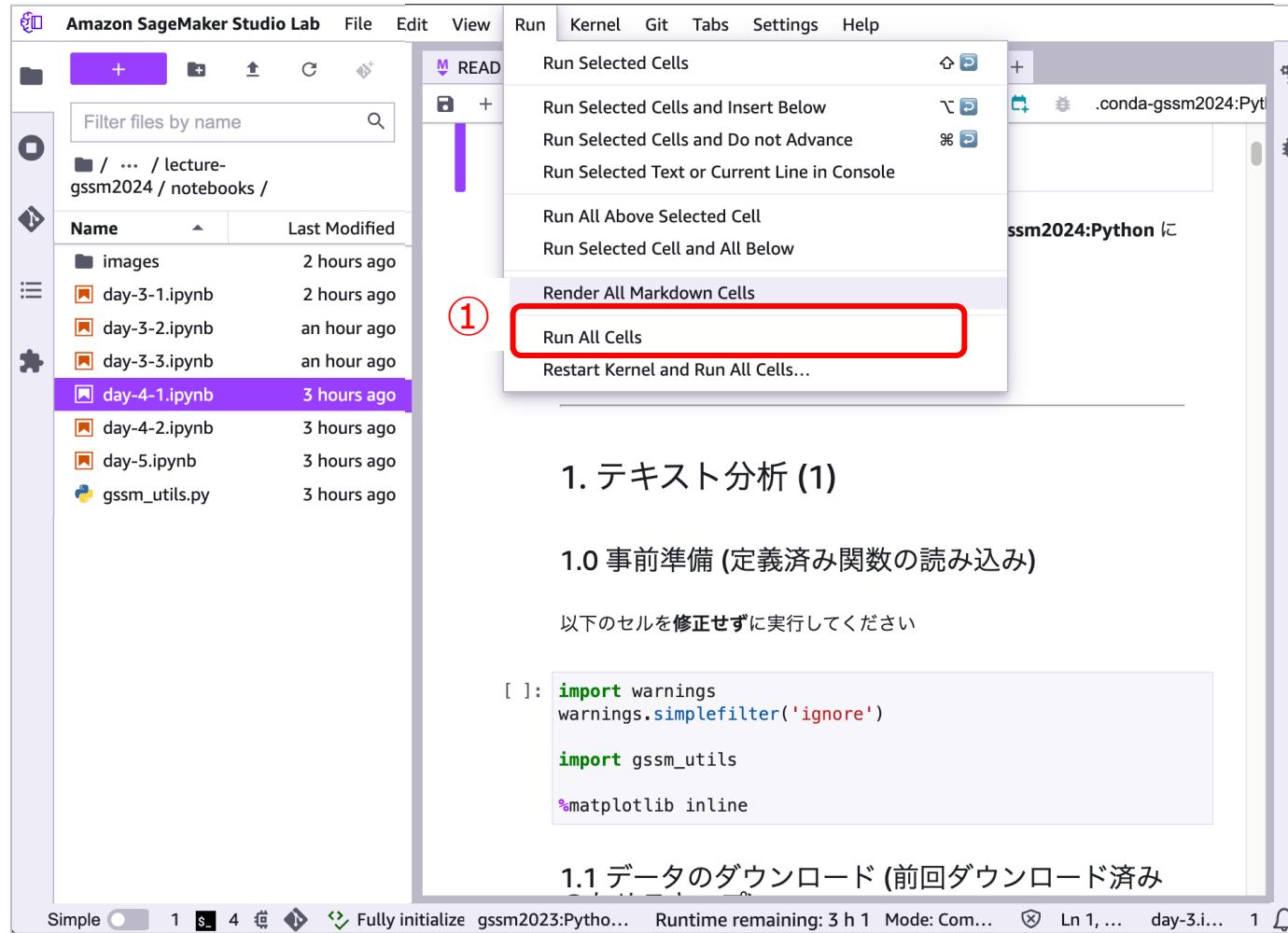
## ● カーネル **.conda-gssm2024:Python** を選択してください **!重要!**

- 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「**.conda-gssm2024:Python**」を選択し、「**Select**」を押す
- ③ 右上隅にカーネル名「**.conda-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-4-1.ipynb" is open. At the top, the "Run" menu is open, showing various options like "Run Selected Cells" and "Run All Cells". A red circle labeled "①" points to the "Run All Cells" option, which is also highlighted with a red box. The notebook content includes sections for "1. テキスト分析 (1)" and "1.0 事前準備 (定義済み関数の読み込み)". Below that is a code cell with imports for `warnings` and `gssm\_utils`, and a magic command `%matplotlib inline`. The bottom status bar shows "Fully initialize gssm2023:Python..." and "Runtime remaining: 3 h 1 Mode: Com...".

演習:

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

この後、Step-by-step で解説します

## テキスト分析 (実践編)

# (再掲) 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	4.25	4.25	4.13	4.05	4.29	4.29	4.30	
01_登別	4.07	4.21	3.95	3.90	4.34	4.08	4.16	
02_草津	4.23	4.22	4.07	3.97	4.32	4.20	4.25	
03_箱根	4.24	4.12	4.18	4.05	4.29	4.33	4.26	
04_道後	4.19	4.41	4.07	4.00	4.03	4.19	4.29	
05_湯布院	4.51	4.28	4.37				4.52	
■ B_ビジネス	3.98	4.30	4.01				4.13	
06_札幌	4.05	4.30	4.09				4.19	
07_名古屋	4.00	4.25	4.04	3.89	3.75		4.15	
08_東京	3.93	4.38	3.94	3.82	3.70	3.99	4.06	
09_大阪	4.01	4.35	4.05	3.93	3.82	4.06	4.18	
10_福岡	3.93	4.24	3.96	3.91	3.64	4.01	4.07	

- 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	4.25	4.25	4.13	4.05	4.29	4.29	4.30
B_ビジネス	3.98	4.30	4.01	3.88	3.74	4.05	4.13

# 実践的な分析

- 実践1: カテゴリーやエリアごとの宿泊者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宿泊者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: ... ....	・... ・...

# 実践的な分析

- 実践1: カテゴリーやエリアごとの宿泊者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宿泊者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: ... • ...	• ... • ...

# 実践1 — 宿泊者の注目ポイントを押さえる

数値評価ではすべての項目に回答  
→ エリアによっても注目する項目にかなり偏りがありそう

## ● 特徴語の抽出結果の例

A_レジャー	数値評価指標
部屋	.328
良い	.318
美味しい	.268
風呂	.265
温泉	.252
スタッフ	.166
ない	.165
宿	.160
露天風呂	.132
夕食	.125

01_登別	02_草津	03_箱根	04_道後	05_湯布院					
温泉	.127	温泉	.149	美味しい	.140	温泉	.106	宿	.164
風呂	.104	風呂	.136	露天風呂	.133	立地	.080	美味しい	.140
食事	.096	宿	.117	風呂	.117	ホテル	.078	露天風呂	.135
サービス	.094	美味しい	.112	部屋	.108	よい	.064	温泉	.113
設備	.085	良い	.107	温泉	.108	浴場	.057	スタッフ	.109
立地	.	.	.	.	.	フロント	.055	風呂	.107

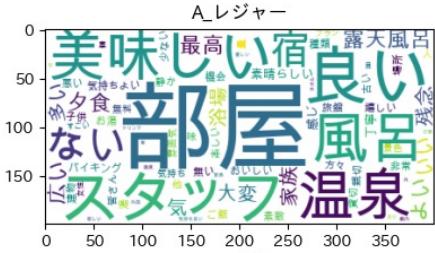
数値評価ではすべての項目に回答  
→ レジャーとビジネスでは注目する項目にかなり偏りがありそう

B_ビジネス	数値評価指標
ホテル	.236
立地	.170
駅	.154
便利	.146
フロント	.116
近い	.112
綺麗	.104
快適	.096
コンビニ	.088
アメニティ	.076

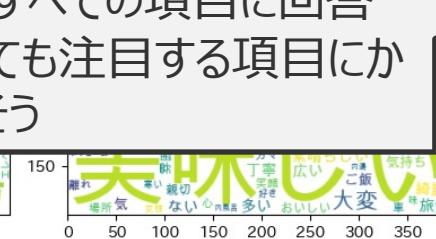
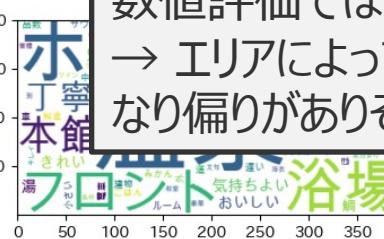
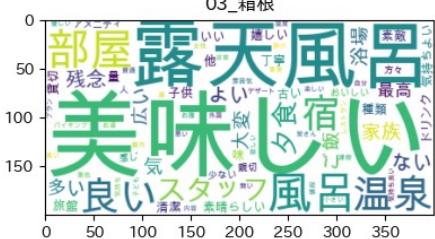
06_札幌	07_名古屋	08_東京	09_大阪	10_福岡					
ホテル	.097	ホテル	.092	駅	.121	駅	.093	立地	.102
部屋	.090	駅	.085	便利	.097	ホテル	.091	ホテル	.099
食事	.	立地	.083	ホテル	.096	部屋	.090	部屋	.091
サービス	.	便利	.072	コンビニ	.082	便利	.084	駅	.085
設備	.	立地	.071	近い	.080	立地	.080	便利	.084
立地	.	近い	.070	立地	.074	フロント	.073	ない	.071
.	.	快適	.066	フロント	.068	近い	.073	綺麗	.067
.	.	広い	.064	駅	.061	アメニティ	.064	フロント	.065
.	.	駅	.062	よい	.061	綺麗	.064	近い	.064
.	.	フロント	.061	アメニティ	.057	快適	.061	バス	.061
.	.	近い	.057	綺麗	.059	快適	.056	コンビニ	.058

# 実践1 — 宿泊者の注目ポイントを押さえる

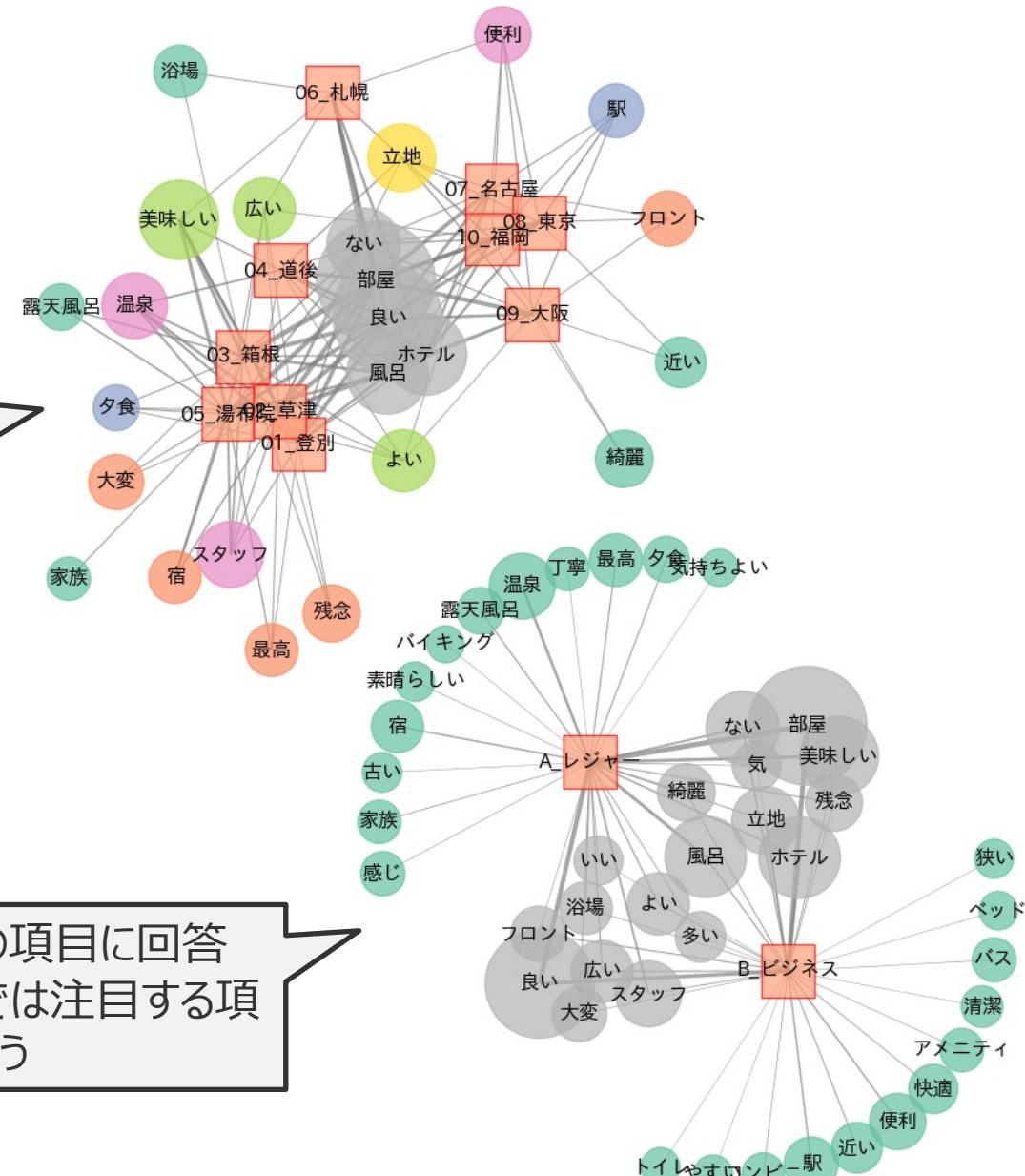
### ● 特徴語の抽出結果の例



数値評価ではすべての項目に回答  
→ エリアによっても注目する項目にかなり偏りがありそう

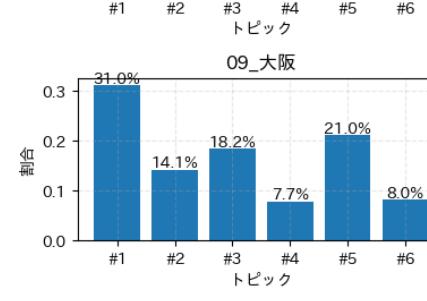
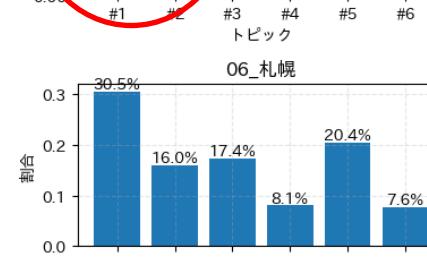
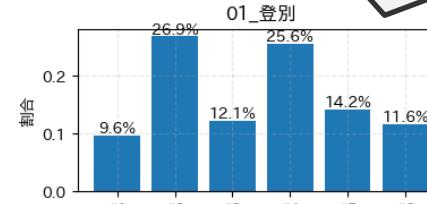
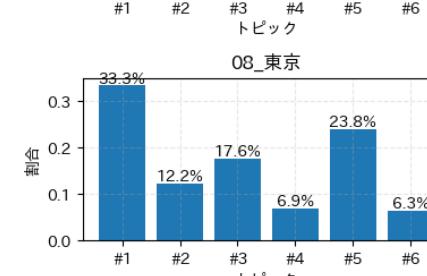
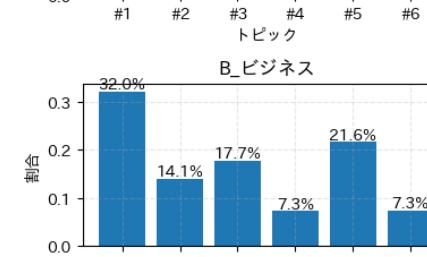
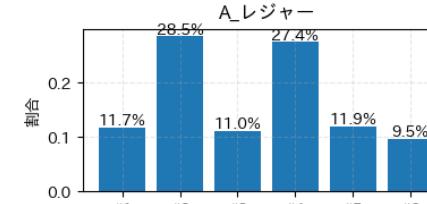
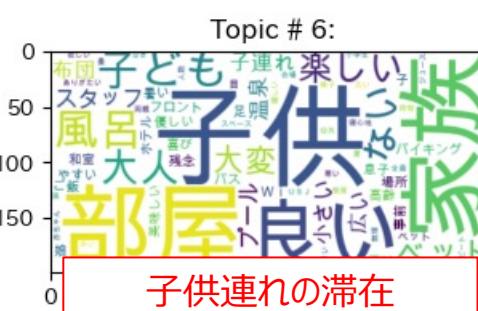
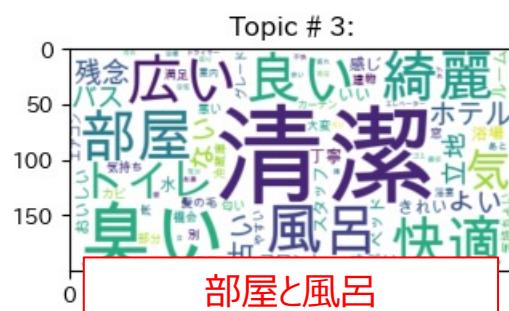
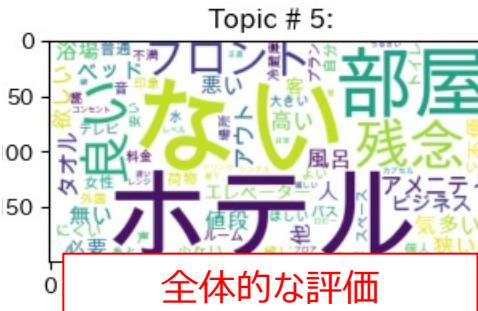
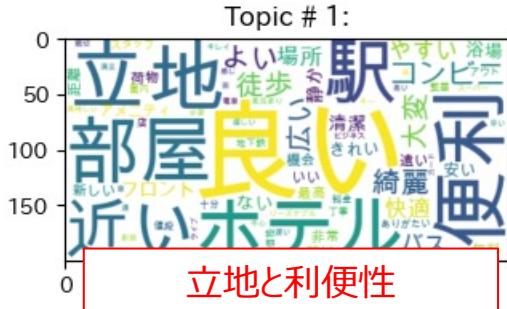


数値評価ではすべての項目に回答  
→ レジャーとビジネスでは注目する項目にかなり偏りがありそう

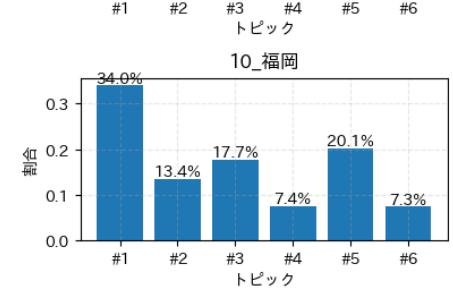
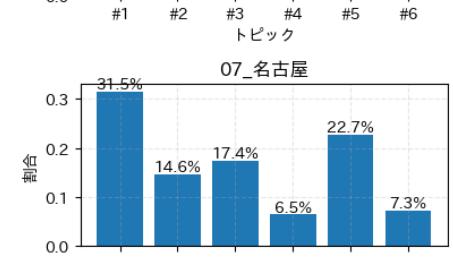
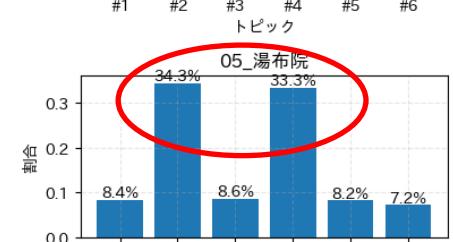
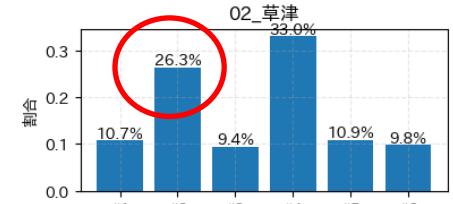


# 実践1 — 宿泊者の注目ポイントを押さえる

## ● トピック抽出結果とトピック割合の可視化例



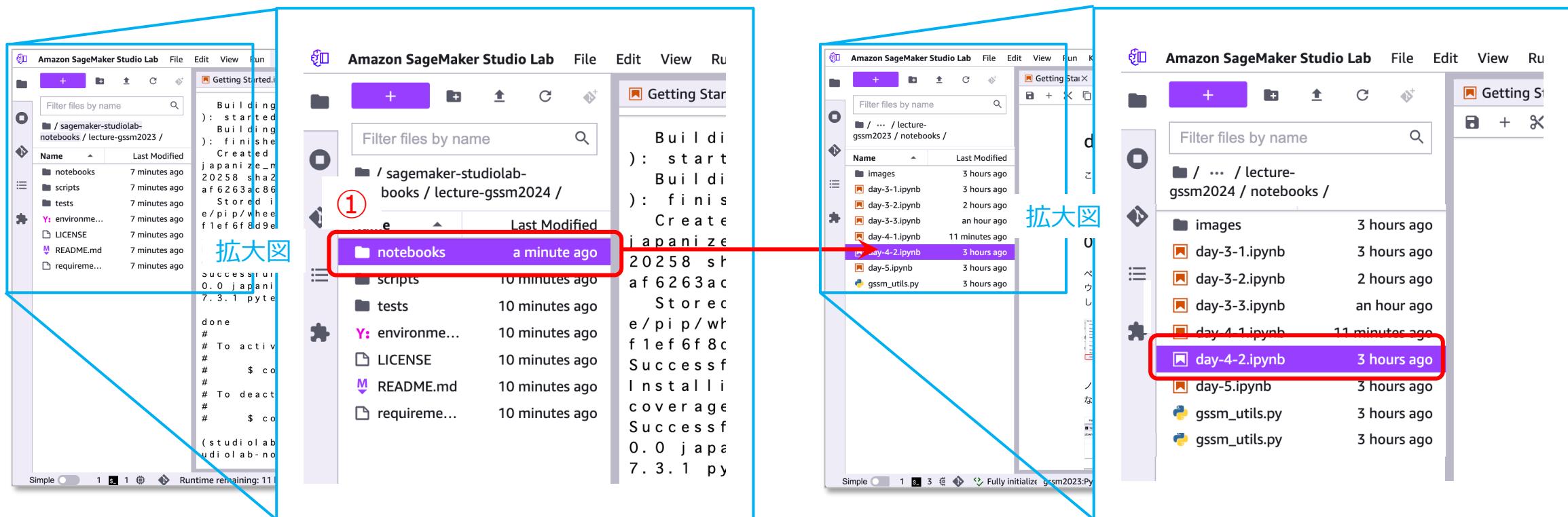
ビジネスに比べてレジャーは、  
地域さが見られる



# 実践1 – 宿泊者の注目ポイントを押さえる

## ● day-4-2.ipynb を開いてください

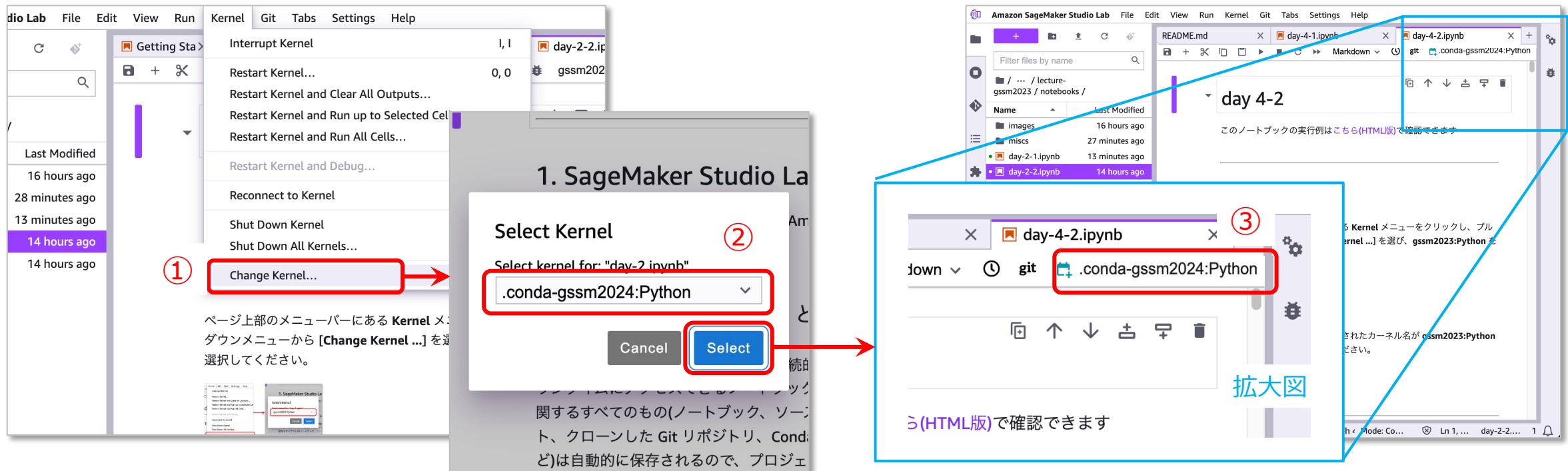
- ① 画面左の **File Browser** から ① **notebooks** をフォルダを開く (既に開いている場合はスキップ)
- ② 次に **day-4-2.ipynb** ノートブックを開く



# 実践1 — 宿泊者の注目ポイントを押さえる

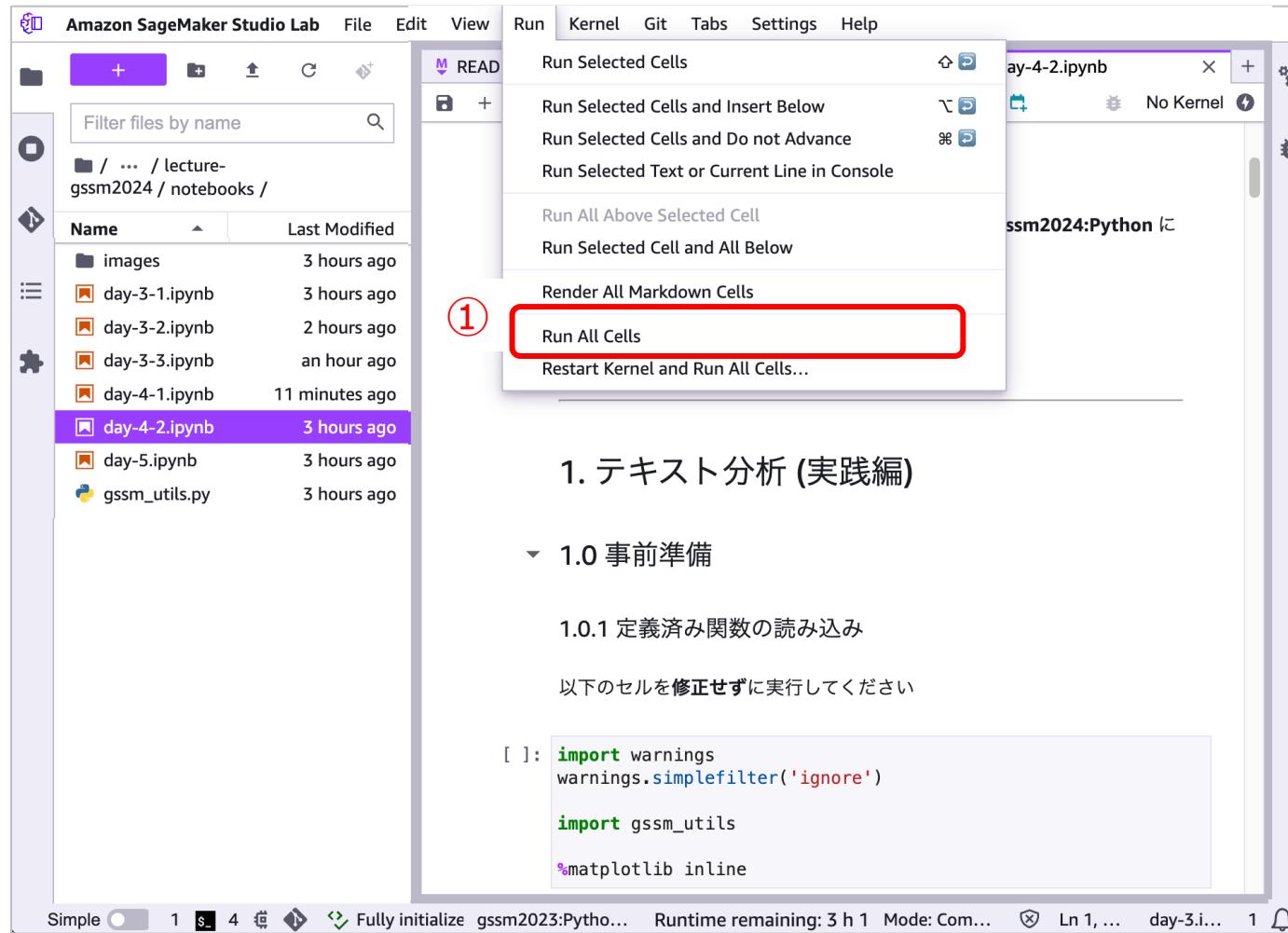
## ● カーネル **.conda-gssm2024:Python** を選択してください **!重要!**

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



# 実践1 — 宿泊者の注目ポイントを押さえる

## ● テキスト解析と可視化



The screenshot shows the Amazon SageMaker Studio Lab interface. On the left, there's a file browser with a list of Jupyter notebooks. In the center, a notebook titled 'ay-4-2.ipynb' is open in a tab labeled 'ssm2024:Python'. The 'Run' menu is open, and the 'Run All Cells' option is highlighted with a red box and a circled '①'. At the bottom, a code cell contains Python code for importing libraries.

```
[ ]: import warnings  
warnings.simplefilter('ignore')  
  
import gssm_utils  
  
%matplotlib inline
```

演習:

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

この後、Step-by-step で解説します

## day 4 – レポート課題

- 以下を PDF ファイルで提出してください
  - ノートブック **day-4-1.ipynb** の末尾にある「【演習】外部変数を利用したエリアごとの作図」に従って作図した 2.1~2.4 の全てのプロットのキャプチャ

※ 何らかの事情で上記のキャプチャ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