

人文社会ビジネス科学学術院 ビジネス科学研究群 2025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 1

- 講義(後半) — 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
- 講義&演習 — データ理解

day 3

- 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)
- 講義&演習 — テキスト分析 (2)

day 5

- テキストマイニングツール紹介 — TMS
- ラップアップ — Q&A

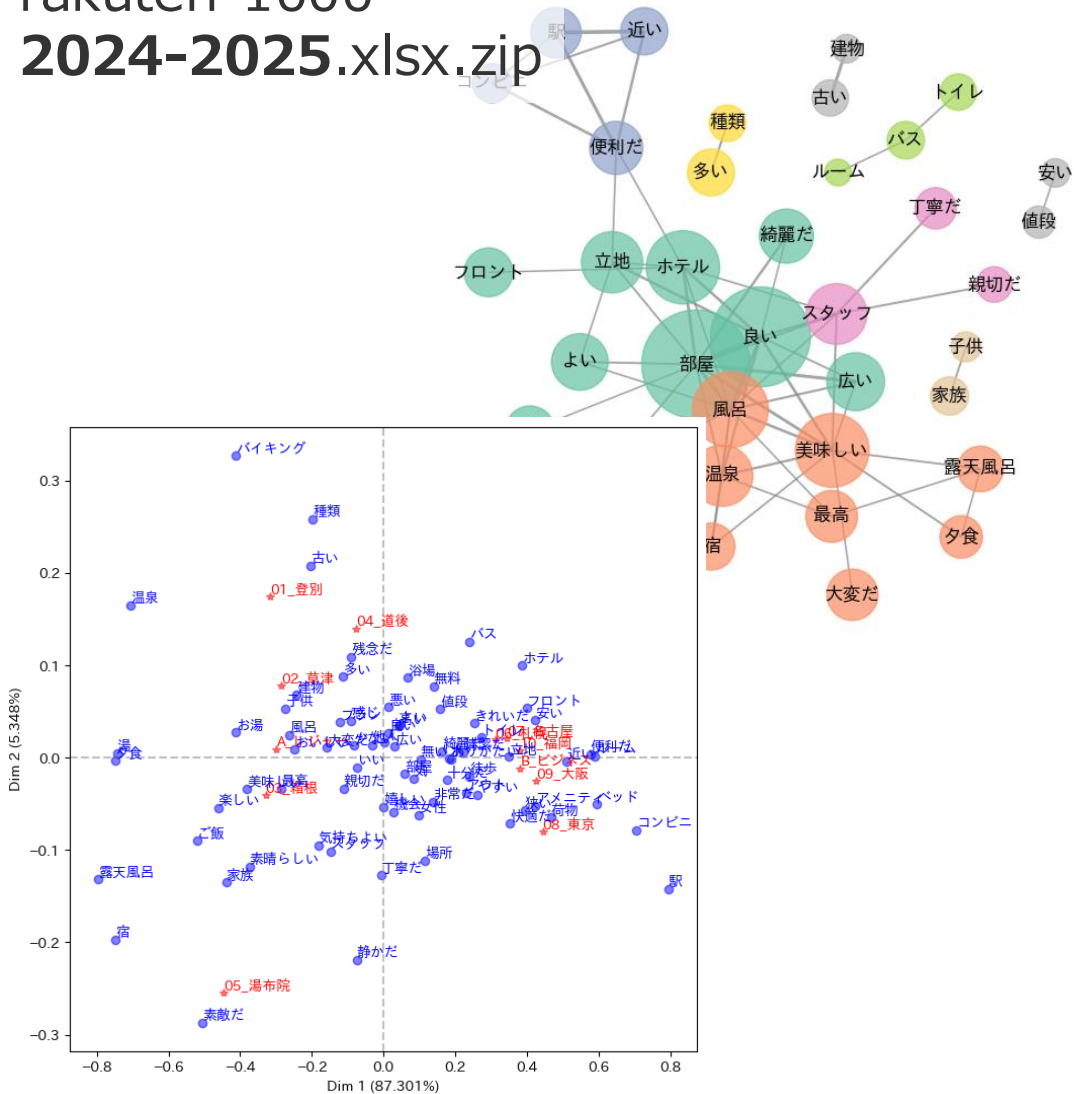
(前回) day 3 – レポート課題

- 以下を PDF ファイルで提出 してください
 - ノートブック **day-3-2.ipynb** の末尾にある「【演習】2022～2023 データセット」に従って、別のデータセット (rakuten-1000-**2022-2023.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

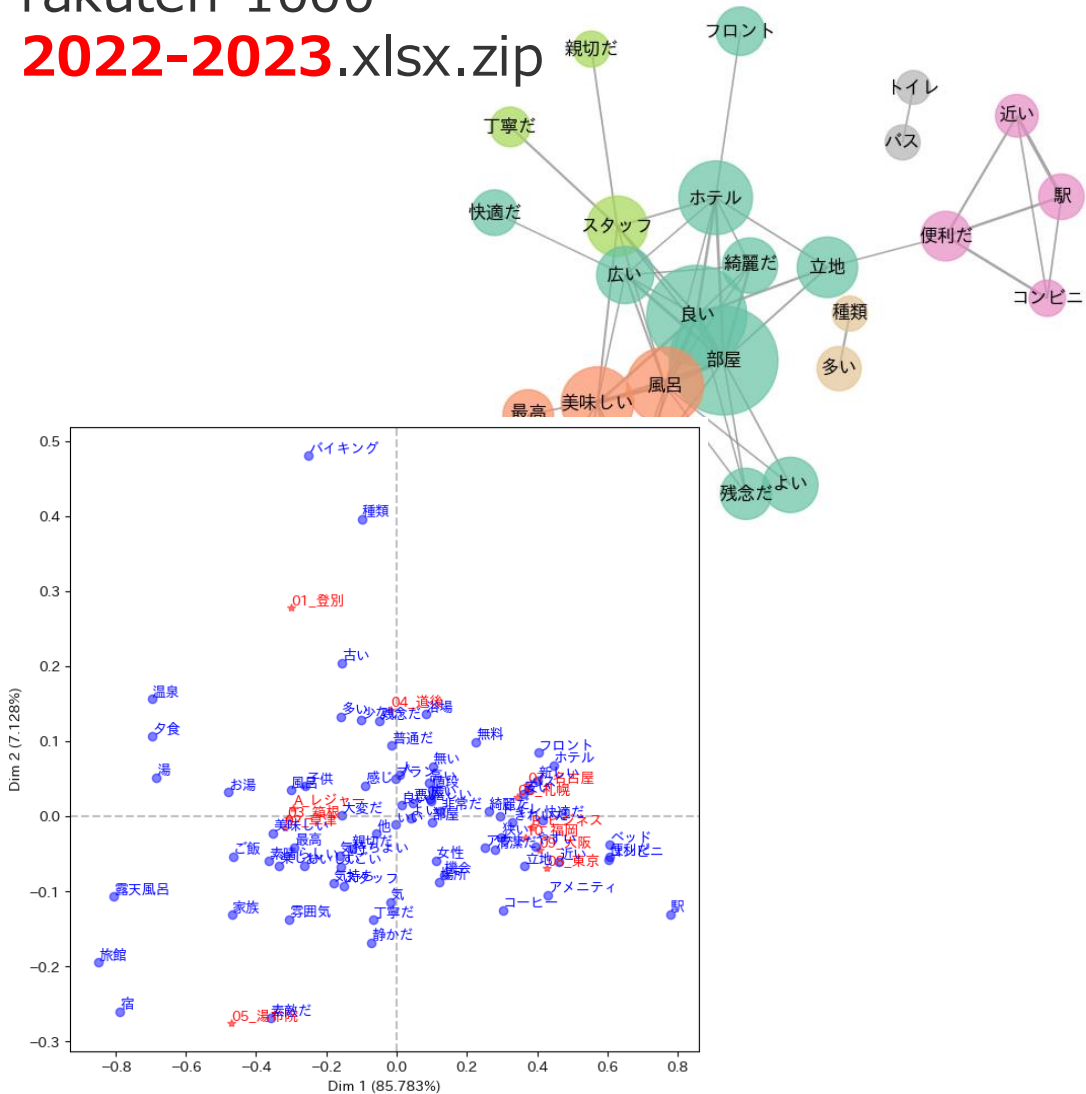
※ 何らかの事情で上記のキャプチャを提出できない場合、本日の講義の感想を文章で記述してください

レポート形式	提出先	期限
PDF	manaba	次回～18:20

rakuten-1000-
2024-2025.xlsx.zip



rakuten-1000-
2022-2023.xlsx.zip



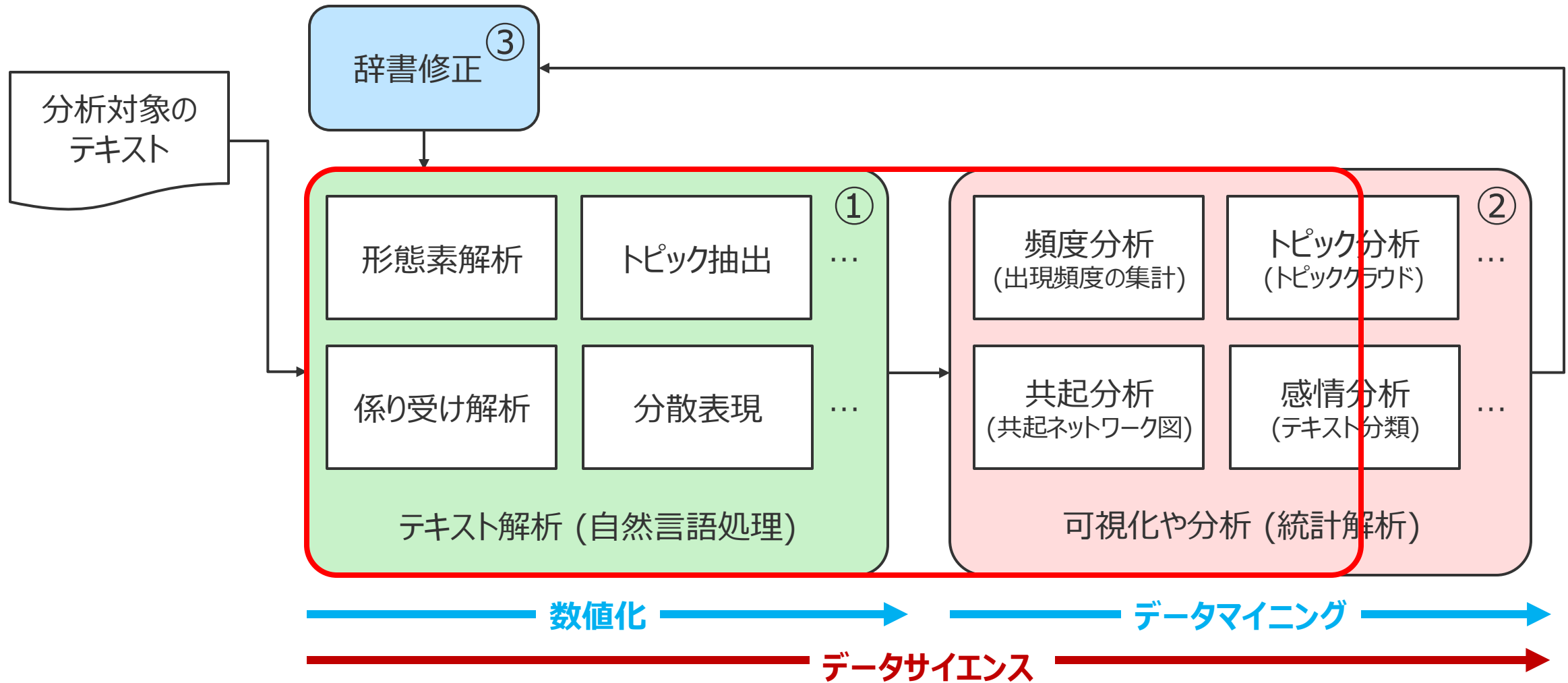
テキスト分析 (1)

(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は？
 - 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(~~無料~~)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: すべて -2010 11 12 13 14 15 16 17 18 19 20 21 22 2023-

著者名: すべて あ か さ た な は ま や ら わ A-Z

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

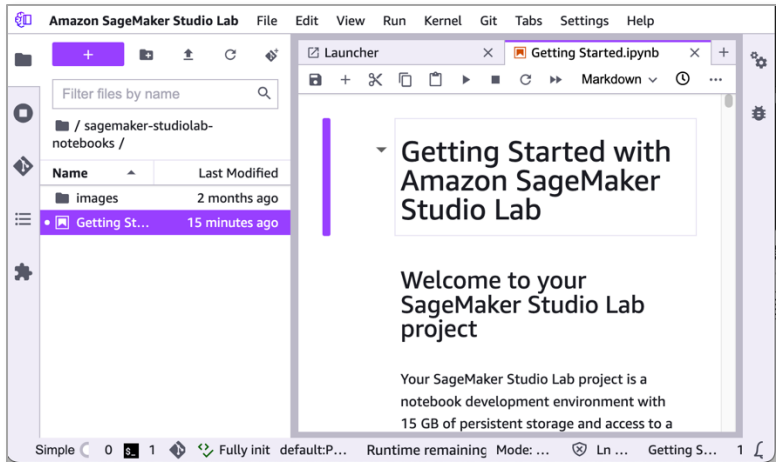
● 機械学習の教育・研究を目的とした研究用ツール



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



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



演習で使用
↓

	Clab(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で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。

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。

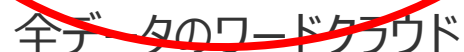
用途: テキスト
プレ

共通する注目ポイントが分かる

直感的
覚的な効

注目ポイントがカテゴリーごとに異なることが分かる

分析例: 宿泊者 注目ポイントを見つける、○○○○

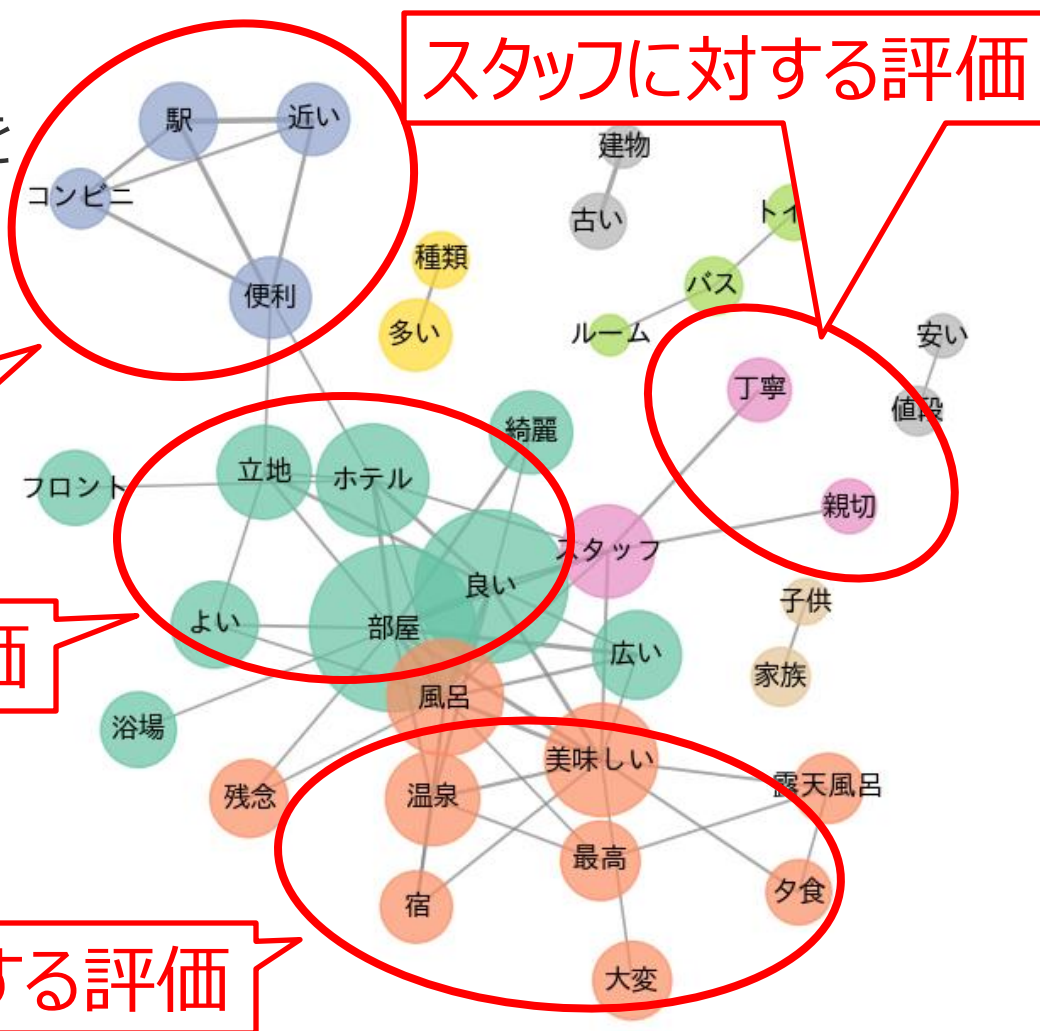


● 共起ネットワーク

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する手法。
ノードが単語をエッジが共起関係を示す。

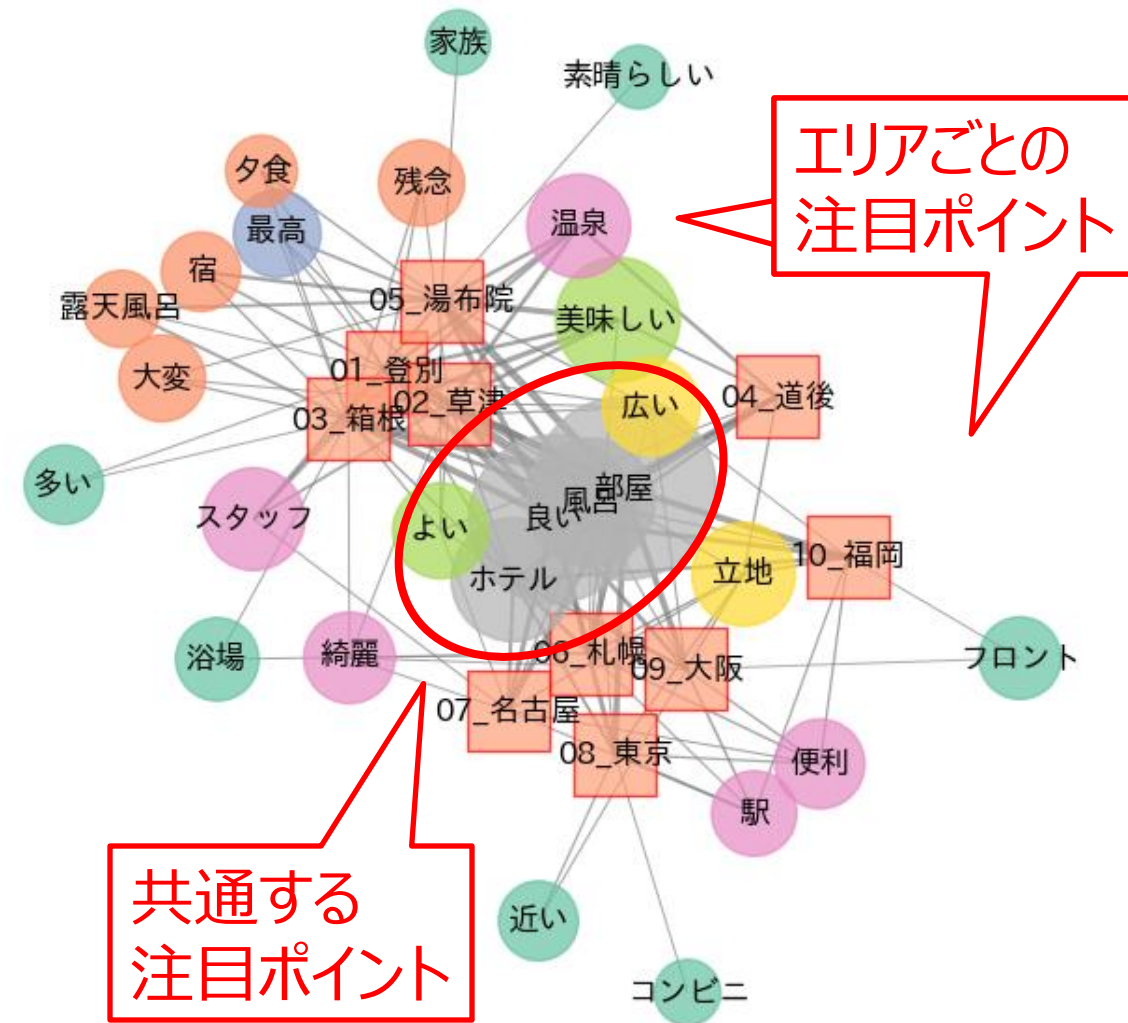
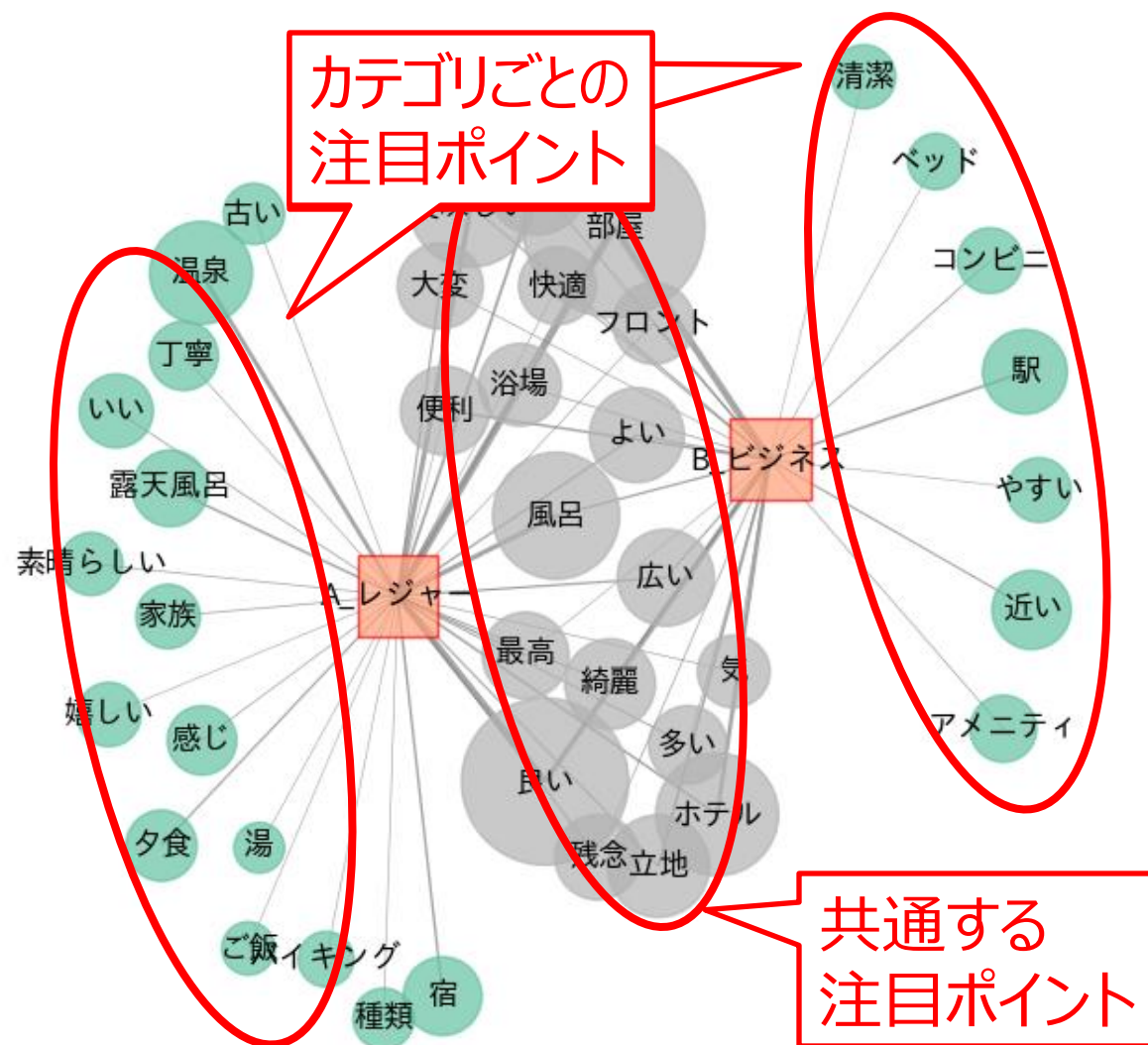
用途: 単語同士の関係性分析するために使用する。
テキスト内のトピックやテーマの繋がりを理解するため

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



全データの係り受けネットワーク図

分析例②: 宿泊者の注目ポイントを見つける、○○ごとに比較する



● 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化する。

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、〇〇ごとに比較する (「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、共起パタンが現れにくい

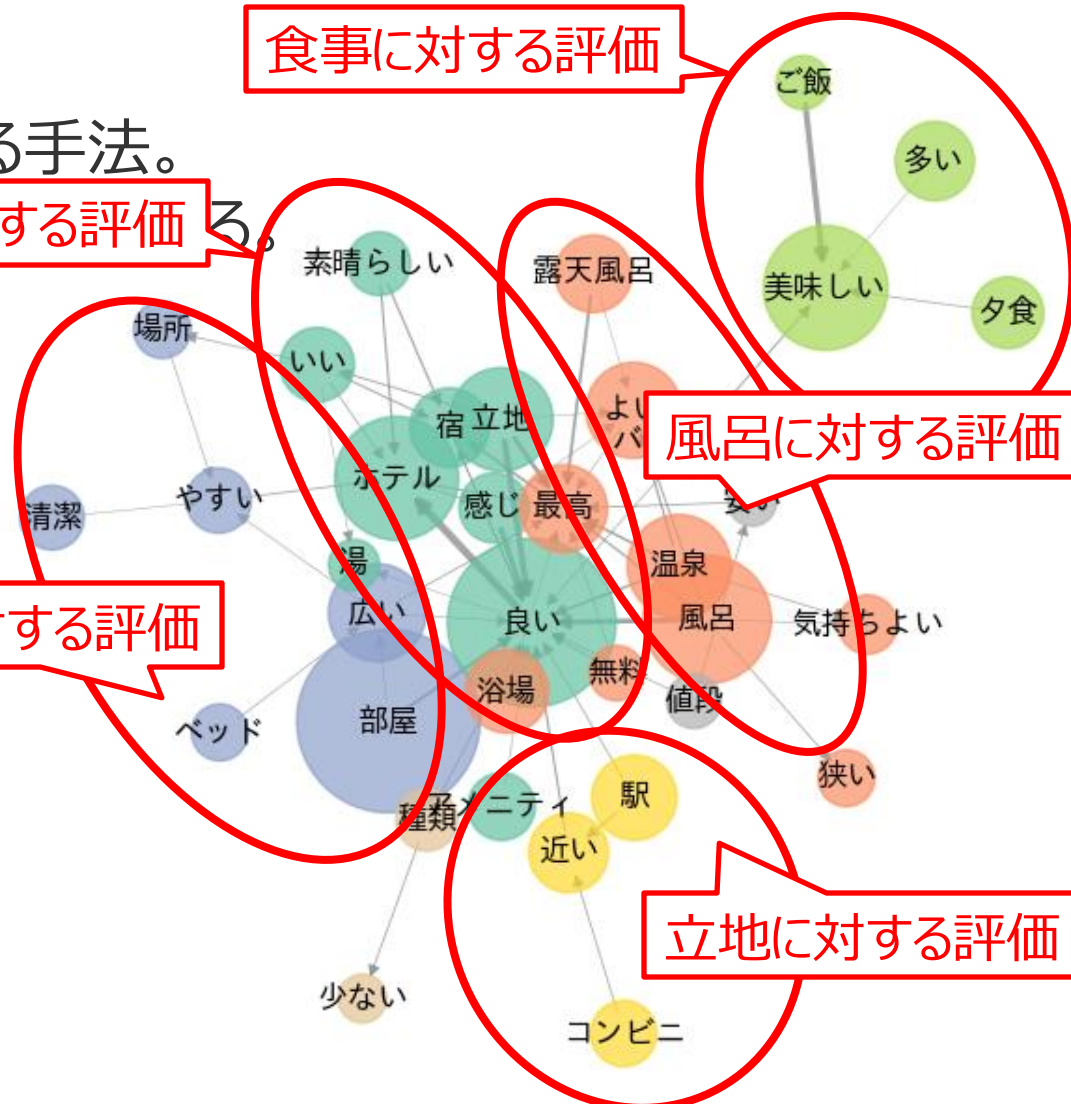
施設に対する評価

食事に対する評価

風呂に対する評価

部屋に対する評価

立地に対する評価



全データの係り受けネットワーク図

● 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、両者の違いを比較する

第2固有値までの累積寄与率は $87.30 + 5.35 = 92.6\%$ で非常に高く、第1,2固有値に対応する軸のみを分析すればよい

横軸は「カテゴリー」で寄与率87%で支配的

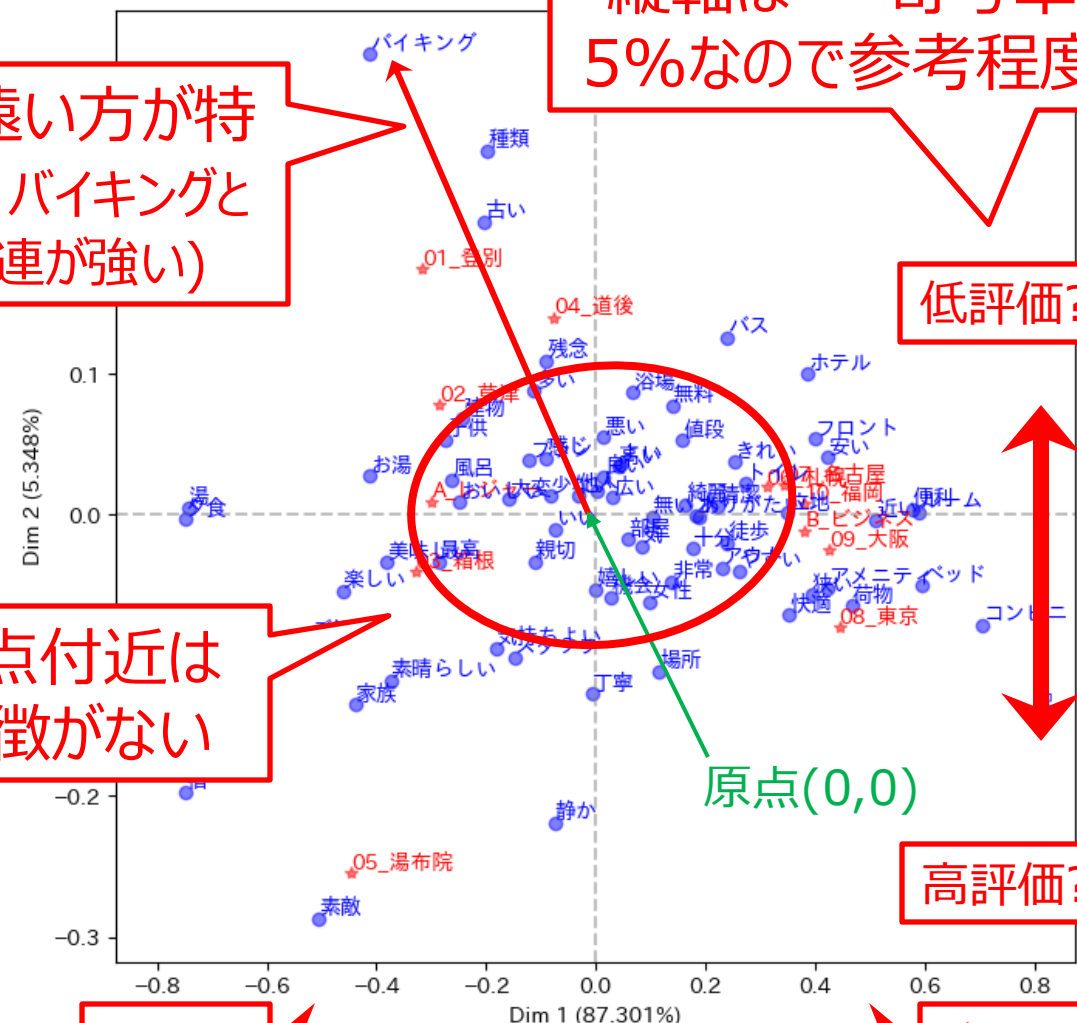
原点から遠い方が特徴的 (例: バイキングと登別は関連が強い)

原点付近は特徴がない

縦軸は...寄与率5%なので参考程度

低評価?

高評価?



● 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相対的な効果も高い

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、○○ごとにトピックの出現割合を比較する

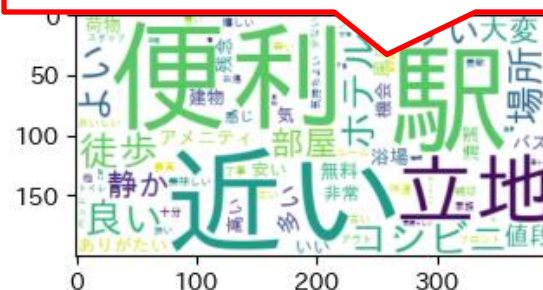
各トピックが含まれる割合も算出できる

温泉・宿泊の満足感



Topic # 4:

立地の便利さとコスパ



Topic # 5:

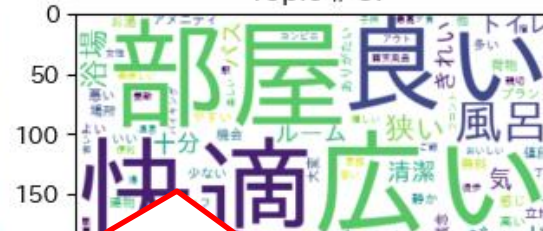
コスパと施設の状態



Topic # 6:



スタッフ対応と接客態度



客室の広さと快適さ



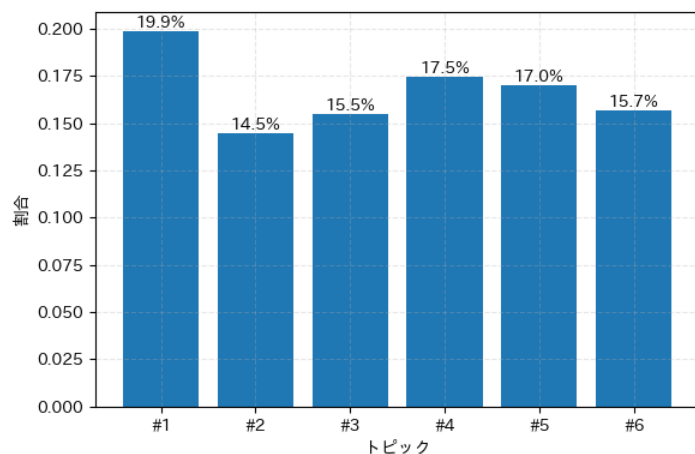
食事と家族向けサービス

● ChatGPT を使ってトピックを説明する

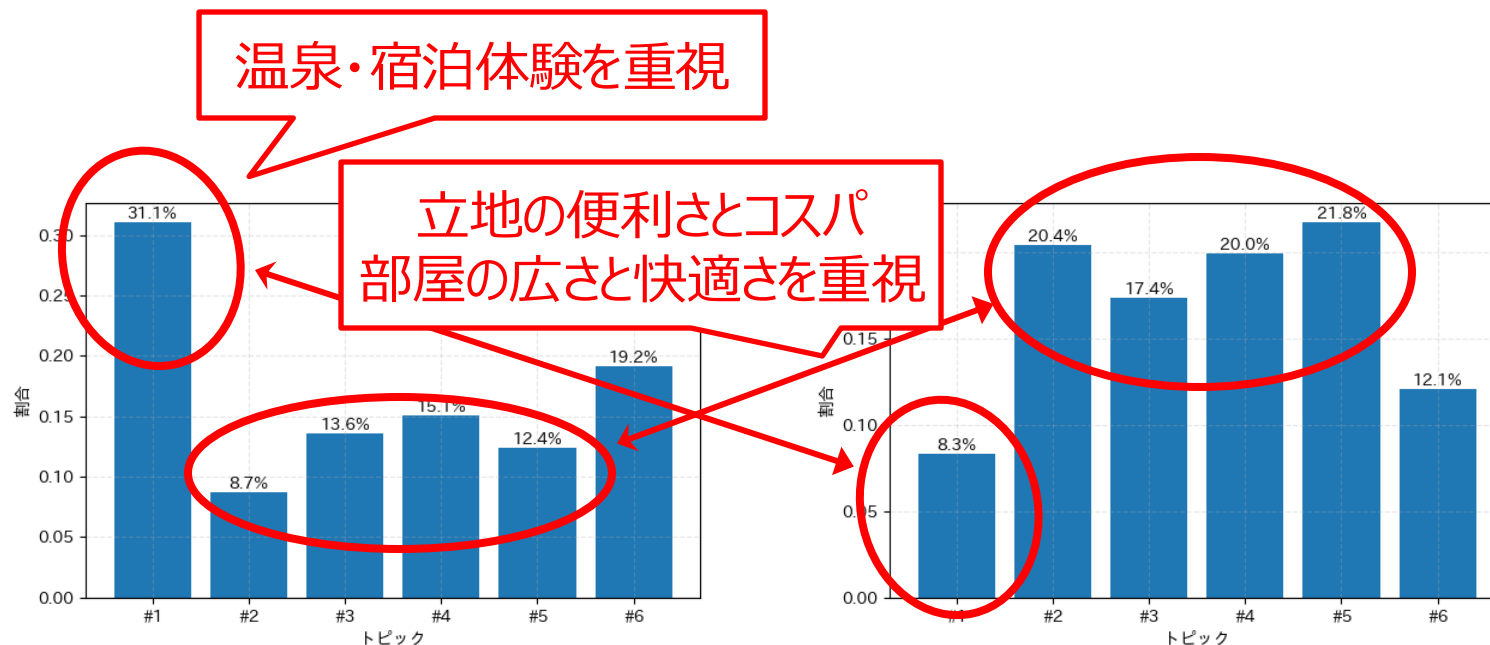
Topic	トピックワード (出現確率 Top20)	トピックの説明 (トピックワードからChatGPTで生成)
# 1	温泉 良い 美味しい 宿 部屋 露天風呂 風呂 スタッフ 大変 夕食 素晴らしい 最高 湯 お湯 楽しい 気持ちよい 素敵 浴場 建物 丁寧	温泉・宿泊体験の満足感 温泉や露天風呂の質の高さ、スタッフの丁寧な対応、美味しい夕食など、宿泊体験全体への高い満足。
# 2	便利 駅 近い 立地 コンビニ 良い ホテル 徒歩 場所 よい やすい 部屋 静か 大変 値段 多い アメニティ 無料 浴場 ありがたい	立地の便利さとコスパ 駅近やコンビニへのアクセスの良さ、安価で便利な点が評価され、アメニティや浴場の無料提供も好印象。
# 3	良い 綺麗 部屋 立地 風呂 ホテル やすい 古い 水 無い 安い 美味しい 残念 高い 感じ 機会 フロント トイレ アメニティ スタッフ	コスパと施設の状態 安さや立地の良さに加え、綺麗さやスタッフ対応の良さもあるが、古さや水回りの不満など、良い点と悪い点が混在。
# 4	ホテル スタッフ フロント 部屋 嬉しい 良い 丁寧 いい アウト 人 他 親切 女性 感じ 悪い 荷物 多い 残念 大変 よい	スタッフ対応と接客態度 フロントやスタッフの親切さ・丁寧さが好印象である一方、対応にばらつきや残念な点も見られる。
# 5	部屋 広い 快適 良い 風呂 ベッド 狭い 浴場 トイレ きれい 清潔 ホテル 気 十分 ルーム バス アメニティ よい 非常 残念	客室の広さと快適さ 部屋の広さや清潔感、設備の充実度が評価されており、快適に過ごせたという声が多いが、一部に狭さや不満の指摘も。
# 6	美味しい 風呂 最高 よい 部屋 種類 バイキング 家族 広い おいしい 子供 多い ご飯 残念 夕食 少ない いい 露天風呂 無料 大変	食事と家族向けサービス バイキングや夕食の美味しさが高評価で、子連れ家族にとって使いやすく、風呂や部屋も広く快適との声が多い。

● トピックモデル (続き)

分析例: 宿泊者の注目ポイントを見つける、○○ごとにトピック出現割合を比較する



全データのトピック構成比率

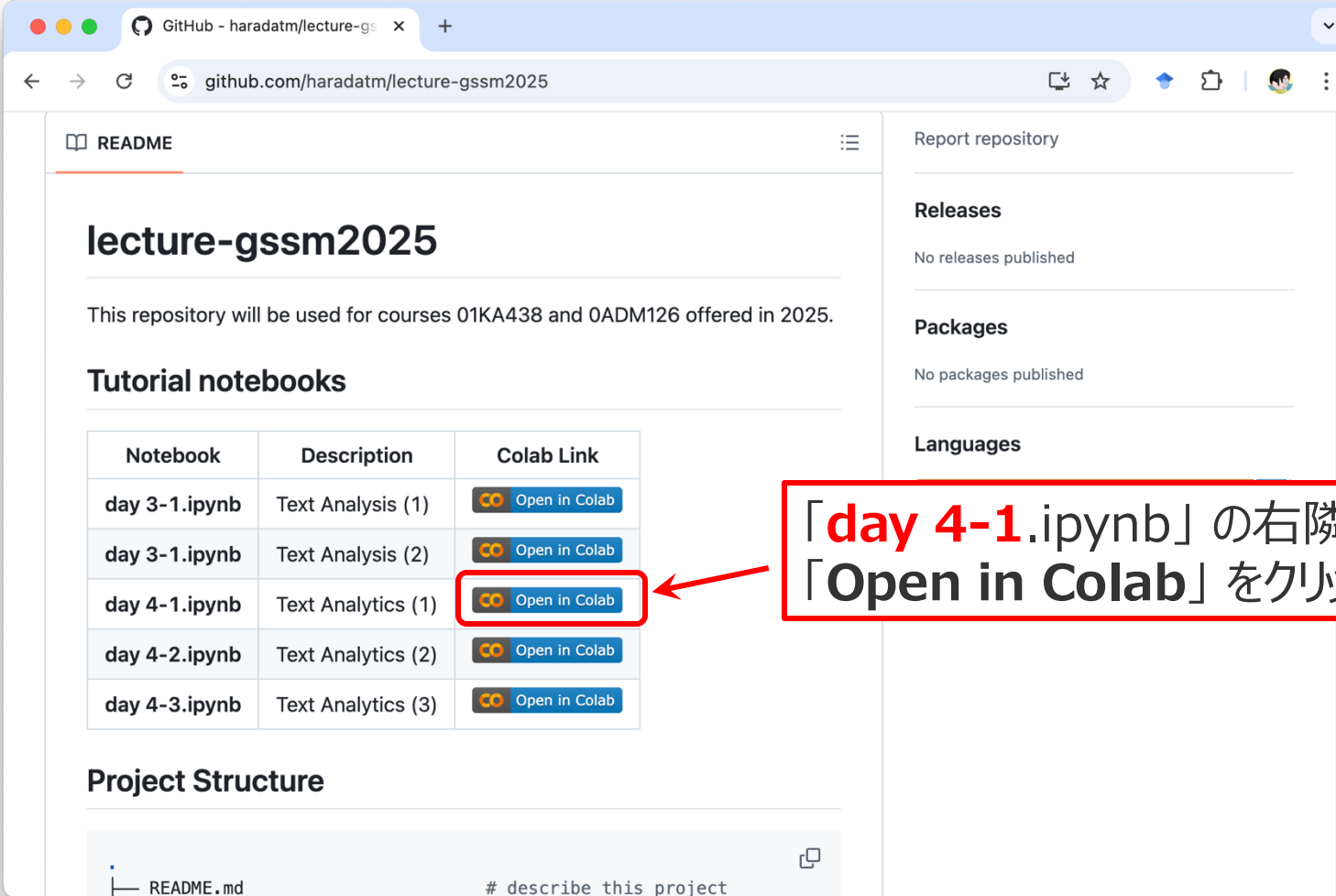


「A_レジャー」のトピック構成比率

「B_ビジネス」のトピック構成比率

「#1 温泉・宿泊体験の満足感」 「#2 立地の便利さとコスパ」 「#3 コスパと施設の状態」
「#4 スタッフ対応と接客態度」 「#5 客室の広さと快適さ」 「#6 食事と家族向けサービスの充実」

- URL: <https://github.com/haradatm/lecture-gssm2025>



The screenshot shows the GitHub repository page for 'haradatm/lecture-gssm2025'. The main content area displays the repository name and a description: 'This repository will be used for courses 01KA438 and 0ADM126 offered in 2025.' Below this is a section titled 'Tutorial notebooks' containing a table with columns 'Notebook', 'Description', and 'Colab Link'. The table lists five notebooks, with the 'day 4-1.ipynb' row highlighted by a red box and an arrow pointing to its 'Open in Colab' link. A red-bordered text box on the right contains Japanese instructions: 「day 4-1.ipynb」の右隣にある「Open in Colab」をクリックして開く. The right sidebar shows sections for 'Report repository', 'Releases', 'Packages', and 'Languages'.

Notebook	Description	Colab Link
day 3-1.ipynb	Text Analysis (1)	Open in Colab
day 3-1.ipynb	Text Analysis (2)	Open in Colab
day 4-1.ipynb	Text Analytics (1)	Open in Colab
day 4-2.ipynb	Text Analytics (2)	Open in Colab
day 4-3.ipynb	Text Analytics (3)	Open in Colab

テキスト分析 (1)

(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x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
01_登別	4.07	4.21	3.95	3.90	4.34
02_草津	4.23	4.22	4.07	3.97	4.32
03_箱根	4.24	4.12	4.18	4.05	4.29
04_道後	4.19	4.41	4.07	4.00	4.03
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
08_東京	3.93	4.38	3.94	3.82	3.70
09_大阪	4.01	4.35	4.05	3.93	3.82
10_福岡	3.93	4.24			4.01

• 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x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: 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

コード: day_4_2_inynb

● 特徴語の抽出結果の例

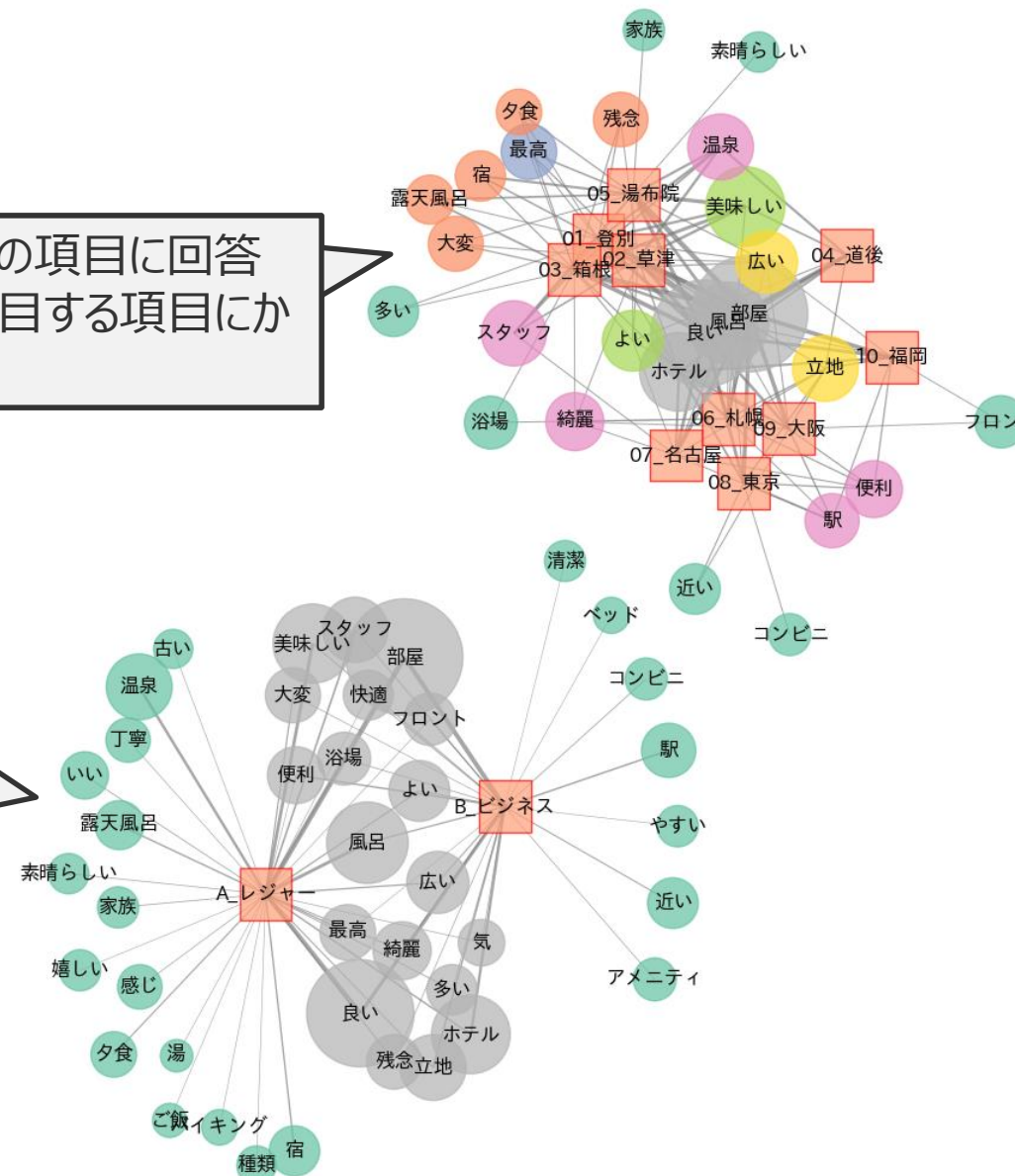
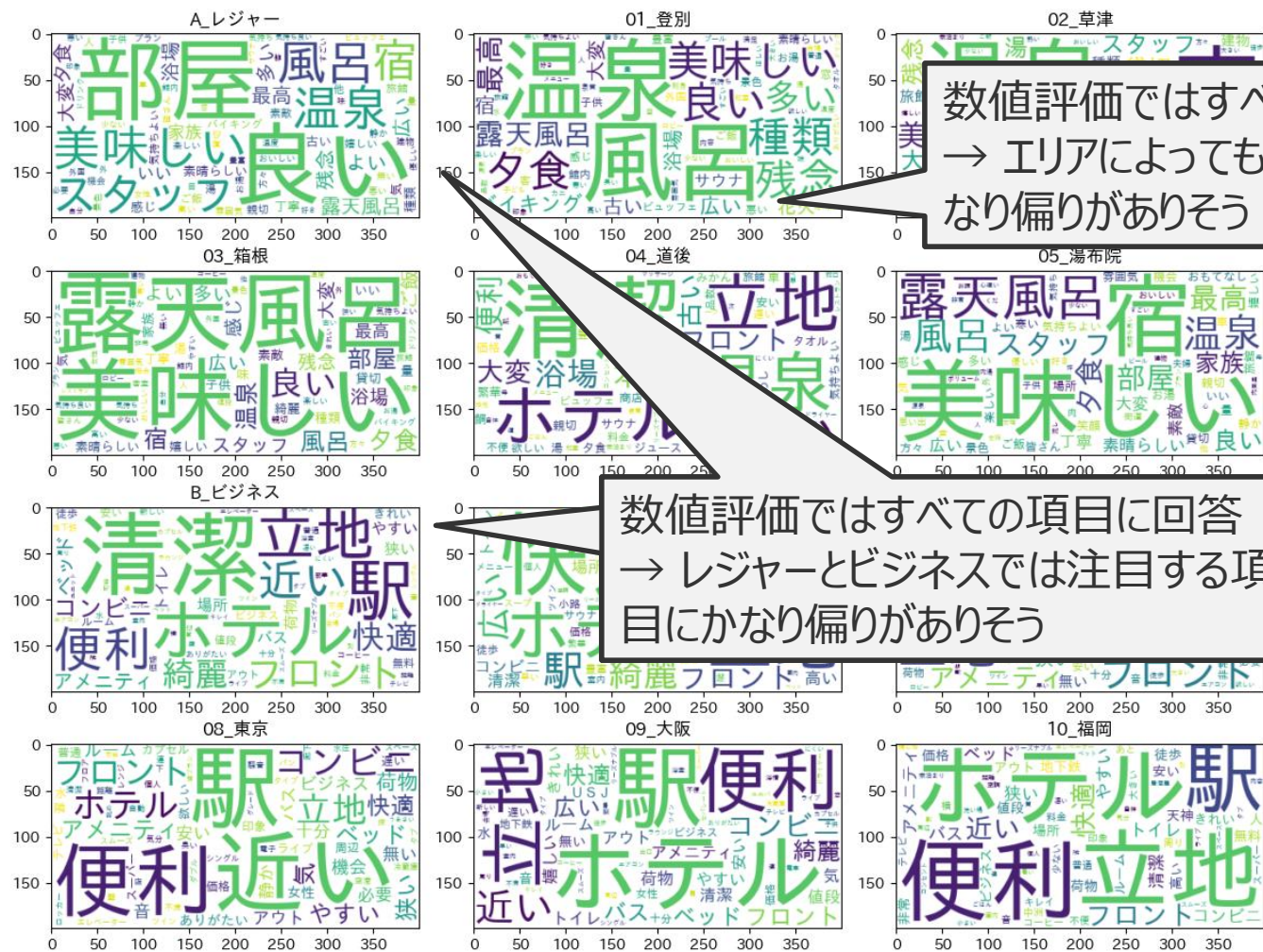
数値評価ではすべての項目に回答
→ エリアによっても注目する項目にかなり偏りがありそう

A_レジャー		数値評価指標		01_登別		02_草津		03_箱根		04_道後		05_湯布院	
部屋	.329	風呂		温泉	.125	温泉	.136	露天風呂	.137	温泉	.120	宿	.165
良い	.309	部屋		風呂	.106	風呂	.130	美味しい	.131	ホテル	.078	美味しい	.149
美味しい	.273	食事		美味しい	.098	宿	.111	良い	.106	立地	.074	露天風呂	.137
風呂	.260	サービス		良い	.092	美味しい	.106	部屋	.105	よい	.066	温泉	.122
温泉	.253	設備		夕食	.085	良い	.103	温泉	.101	浴場	.061	風呂	.119
スタッフ	.155	立地							.101	本館	.057	スタッフ	.116
宿	.148								.097	便利	.055	最高	.107
露天風呂	.145								.091	フロント	.054	部屋	.107
最高	.131								.090	大変	.052	夕食	.105
夕食	.123								.086	いい	.047	家族	.095

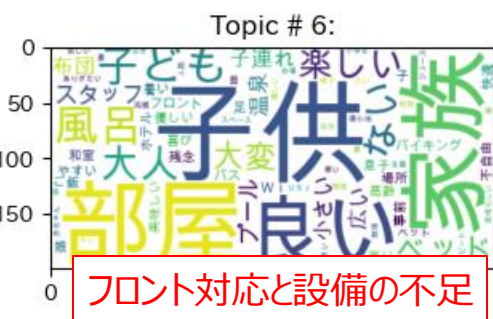
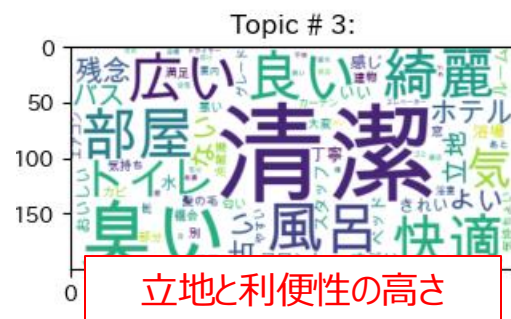
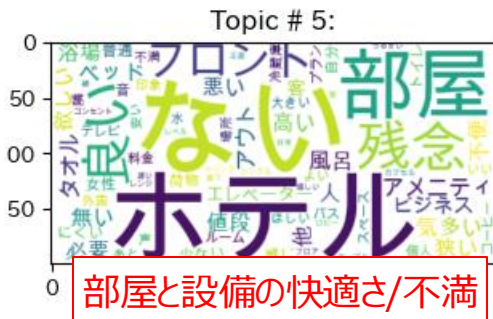
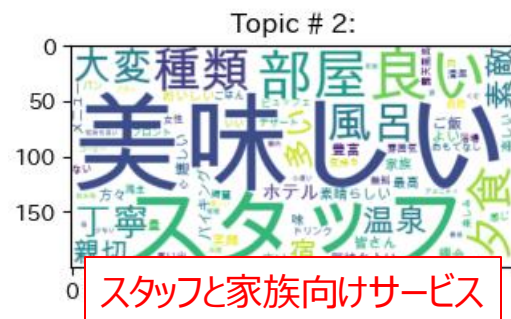
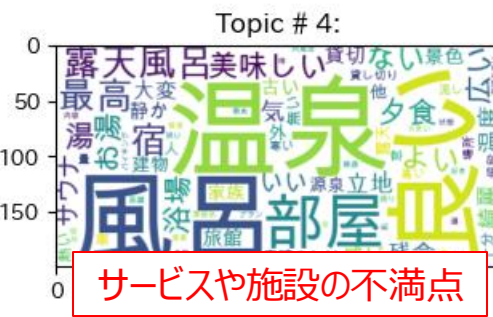
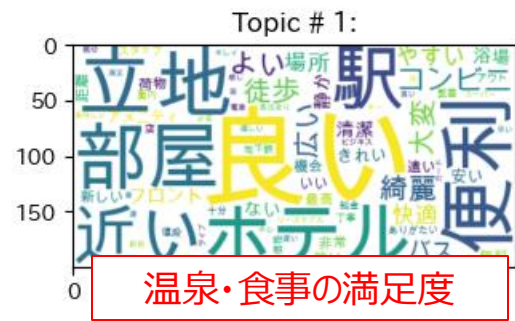
数値評価ではすべての項目に回答
→ レジャーとビジネスでは注目する項目にかなり偏りがありそう

B_ビジネス		数値評価指標		06_札幌		07_名古屋		08_東京		09_大阪		10_福岡	
ホテル	.216	風呂		ホテル	.096	ホテル	.088	駅	.123	駅	.099	ホテル	.095
駅	.153	部屋		便利	.083	駅	.070	近い	.083	ホテル	.092	立地	.085
立地	.152	食事		立地	.076	便利	.069	便利	.082	便利	.080	便利	.082
便利	.142	サービス		浴場	.070	立地	.064	コンビニ	.081	立地	.077	駅	.080
近い	.108	設備		駅	.070	フロント	.064	ホテル	.081	近い	.075	近い	.066
フロント	.105	立地		快適	.063	アメニティ	.061	立地	.076	コンビニ	.069	フロント	.066
綺麗	.103			綺麗	.062	綺麗	.059	フロント	.061	フロント	.065	快適	.056
快適	.093			広い	.059	快適	.055	アメニティ	.056	快適	.063	コンビニ	.051
コンビニ	.087			フロント	.058	近い	.055	快適	.054	広い	.063	アメニティ	.051
アメニティ	.074			コンビニ	.054	コンビニ	.047	やすい	.052	綺麗	.059	ベッド	.045

● 特徴語の抽出結果の例



● トピック抽出結果とトピック割合の可視化例



サービスや施設を重視

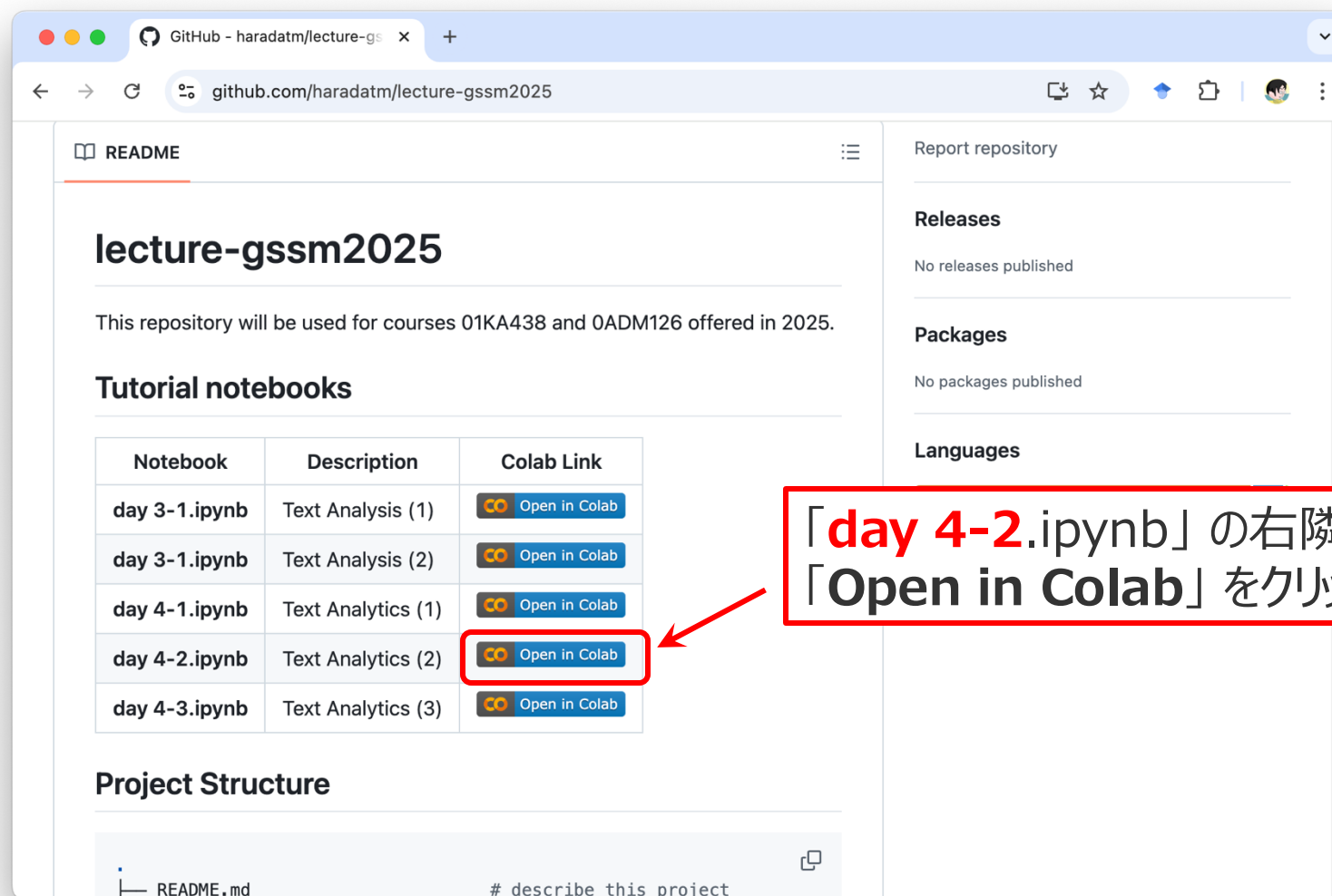
立地と利便性を重視

温泉・食事を重視

ビジネスに比べてレジャーは、地域さが見られる



- URL: <https://github.com/haradatm/lecture-gssm2025>



テキスト分析 (2)

(再掲) 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: ... •...	•... •...