

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 2

スケジュール

day 1

- 講義(後半) – 自然言語処理とLLM

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
- 講義&演習 – テキスト解析 (1)

day 3

- 講義&演習 – テキスト解析 (2)
- 講義&演習 – テキスト分析 (1)

day 4

- 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 1 – レポート課題

- 以下を PDF ファイルで提出 してください
 - 社会人学生に役立つ、お薦めの生成AI(LLMやテキストマイニング以外も可)と使い方を教えてください
- ※ 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

参考－生成AI活用アイデア(7/10時点の21件分で集計)

業務効率化・自動化	<ul style="list-style-type: none"> 議事録の自動作成・要約: 会議の音声データから自動で文字起こしを行い、要点をまとめた議事録を作成する 問い合わせ対応の自動化: FAQやマニュアルを学習させ、社内外からの問い合わせに自動で回答するチャットボットを構築する 文章の校正・翻訳: 作成した文章の誤字脱字チェックや、より適切な表現への修正、多言語への翻訳を瞬時に行う 社内文書の検索と要約: 社内の膨大な規定やマニュアルの中から、必要な情報を探し出し、要約して提示させる 採用業務の効率化: 応募者のエントリーシートをAIが分析・評価し、面接官の負担を軽減する
コンテンツ制作	<ul style="list-style-type: none"> マーケティングコンテンツの生成: 広告のキャッチコピー、SNS投稿文、メルマガなど、ターゲットに合わせた文章を複数パターン生成する 資料作成の支援: プрезентーションの構成案作成、スライドの文章やデザイン案の生成をサポートする アイデア創出(ブレインストーミング): 新規事業や企画のアイデア出しの壁打ち相手として活用し、多様な視点を得る デザイン・画像生成: ロゴ、バナー、イラストなど、コンセプトを伝えるだけでデザイン案を複数作成させる
専門業務の支援	<ul style="list-style-type: none"> ソフトウェア開発: プログラムコードの自動生成、レビュー、デバッグ作業を支援させ、開発スピードを向上させる 教育・研修: 個人の理解度や職種に合わせた研修コンテンツやテスト問題を自動生成し、個別最適な学習を提供する 法務・契約業務: 契約書のドラフト作成や、不利な条項がないかのリーガルチェックを支援させる

1. 文章生成・要約・翻訳

ツール・サービス名	登場回数
ChatGPT	6
Gemini	3
Copilot	2
Claude	2
NotebookLM	2
自治体 AI zevo	1
Readable	1
Cedric	1
DeepL	1
Google翻訳	1
Grammerly	1

2. 情報収集・調査

ツール・サービス名	登場回数
Consensus	1
Connected Papers	1
Scholar AI	1
Perplexity	1

4. プログラミング・開発

ツール・サービス名	登場回数
Cursor	2
Mastra AI	1
OpenAI BatchAPI	1

6. その他

ツール・サービス名	登場回数
Slack AI	1
Box AI	1
Google Drive OCR	1

3. 資料作成・プレゼンテーション

ツール・サービス名	登場回数
Mapify	3
Canva AI	1
イルシル	1
Gamma	1

5. 音声・動画処理

ツール・サービス名	登場回数
Gladia	1
Otter	1

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

クチコミサイトの例 — 楽天トラベル

● ホテルのクチコミ数: 1,468万件 ※年間約60~80万

The screenshot shows the Rakuten Travel website at the URL <https://travel.rakuten.co.jp/review/>. The main heading is 'お客様の声' (Customer Reviews) with the number '14,648,306'. Below it, there's a search bar for reviews and filters for domestic and overseas hotels. A sidebar displays the latest reviews, including one for 'ダイワロイネットホテル那覇国際通り' (May 23, 2025) and another for 'プレミアホテルーCABIN PRESIDENTー函館' (May 23, 2025). A callout box highlights that the site has over 14 million reviews.

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 (2023)
 - 1,393万件 (2024)
 - **1,468万件 (今回)**
- ※ 2025/5/24現在

鶴川シーワールドホテル クチコミ

HARADA Tomohiko

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

検索 サーチ

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](#)

楽天トラベル
スーパーDEAL
30%以上ポイントバック!

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・お客さまの声(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 [Facebook] [Twitter] [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

地図

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
- ★★トッページ★
- 鶴川シーワールドホテル
- ★鶴シニース★

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

施設紹介 プラン一覧 フォトギャラリー(76) 地図・アクセス お客さまの声(886) クーポン一覧 プレゼント

鶴川シーワールドホテルのクチコミ・お客さまの声

●ホテル・旅行のクチコミTOP

総合評価		評価内訳		項目別の評価	
★★★★★	4.12	アンケート件数: 886件	5点	★★★★★ 4.11	サービス
			4点	★★★★★ 4.61	立地
			3点	★★★★★ 3.53	部屋
			2点	★★★★★ 3.62	設備・アメニティ
			1点	★★★★★ 3.53	風呂
				★★★★★ 4.10	食事

投稿の種類	宿泊年月	キーワード
<input checked="" type="checkbox"/> クチコミ (感想・情報) <input type="checkbox"/> 同伴者	指定なし (7)	□
<input type="checkbox"/> 一人 <input type="checkbox"/> 家族 <input type="checkbox"/> 恋人 <input type="checkbox"/> 友達 <input type="checkbox"/> 仕事仲間	年代	□
	指定なし	□
	性別	□
	□	取り込みを解除する

[並びかえ] 最新の投稿順 評価が高い順 (総合 | サービス | 立地 | 部屋 | 設備・アメニティ | 風呂 | 食事)

772件中 1~20件表示 [1 | 2 | 3 | 4 | 5 | ... 全 39 ページ] ●次の20件

総合 ★★★★★ 4
RENDEZ_VOUSさんの 鶴川シーワールドホテル のクチコミ (感想・情報)

RENDÉZ_VOUSさん (3件) [30代/女性] 2015年06月17日 19:20:02

入園パスポート付き、バイキングの夕食と朝食付きでとてもお得な価格で泊まれました。

2日間ともシーワールドに入園出来るのでとても便利です。バイキングも種類が豊富でおいしく楽しくいただきました。温泉は湯船がひとつしかないのが少し物足りないですが、メインはシーワールドなので仕方ないかなと。

お部屋は古くて、他の部屋の物音(子供が走り回る足音など)が気になりました。

量の上に直に座ったらあちこちかゆくなりました。

オーシャンビューで部屋の目の前が海ではうれしいのですが、波の音が大きくて聞こえて、なかなか眠れませんでした。

設備が古いから仕方ないのかな。

総評すると 部屋はちょっと微妙ですがお安いですし満足しています。

シーワールドのシャチのショーを金曜日と土曜日に見ましたが、全然レベルが違かった。(笑)

平日は練習程度のレベルでした。土曜日にもう一度見てよかったです。

シャチ以外のショーは平日でもしっかりしていました。

レビューを評価してください
このレビューは参考になりましたか?

旅行の目的 … レジャー
同伴者 … 家族
宿泊年月 … 2015年06月

不適切なレビューを報告する
いいえ いいえ

ご利用の宿泊 いい値!バリュープラン
プラン ご利用のお部屋 [wa海側和室 (10畳バス・トイレ付) タイプ]

● クチコミを投稿する
* クチコミを修正する

宿泊プラン一覧

[1泊朝食付 & カミソリ入園バス付] 朝からカモシーカ Let's Go!
[最安料金 (国安)] 7,963円~
(消費税込6,600円~)

[1日バスポート+朝食付 +翌日のランチ券付] 週くわいくてたっぷり楽しめる!
[最安料金 (国安)] 8,519円~
(消費税込9,200円~)

[1泊朝食付] カモシーカ
[最安料金 (国安)] 9,352円~
(消費税込10,100円~)

[カモシーオリジナルポンチョ付] 6月のお得Days
[最安料金 (国安)] 9,538円~
(消費税込10,300円~)

[シャチのイラスト入りオリジナルフェイスタオル付] 7月のお得Days
[最安料金 (国安)] 9,815円~
(消費税込10,600円~)

[3日前までの予約] 早めに決めてお得♪プラン
[最安料金 (国安)] 10,000円~
[モーニングブレックファスト付]

R 鶴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★
●鶴川シーワールドホテル
★★レストラン★
●鶴川シーワールドホテル
★★温泉大浴場★
●鶴川シーワールドホテル
★★施設・設備★
●鶴川シーワールドホテル
★★よくあるご質問★
●鶴川シーワールドホテル
★★アクセス★
●鶴川シーワールドホテル
設備・アメニティ・基本情報
●鶴川シーワールドホテル
写真・画像
●鶴川シーワールドホテル
地図・アクセス
●鶴川シーワールドホテル
クチコミ
●鶴川シーワールドホテル
温泉

外国语サイト
●Book Kamogawa Sea World Hotel
●Hotels in Sotobo(Kamogawa, Katsura, Onjuku, Mabora)
●KAMOGAWA SEA WORLD HOTEL 預訂
●外房(鶴川・勝浦・御宿・茂原)酒店一覧

★★★★★ 2
投稿者さんの 鶴川シーワールドホテル のクチコミ (感想・情報)
<p> 投稿者さん 2015年06月11日 17:03:57</p> <p>良かったところ</p> <ul style="list-style-type: none"> ・部屋からの景色（朝日最高でした） ・食事（品数が多く、朝ごとも良かったです） ・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。 <p>掃除が行き届いているとの口コミを多く見ましたが、そうは思いませんでした。</p> <p>気にかかる事は多々ありました、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。</p>
<p> 鶴川シーワールドホテル 2015年06月11日 19:32:50</p> <p>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。</p> <p>客室内清掃の件、大変申し訳ございませんでした。</p> <p>重要改善として、早急に対応いたします。</p> <p>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。</p> <p>フロントスタッフへのお言葉、誠にありがとうございました。</p> <p>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、</p> <p>スタッフと共にさせて頂きます。</p>
<p>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?</p> <p>いいえ いいえ</p> <p>旅行の目的 … レジャー</p> <p>同伴者 … 家族</p> <p>宿泊年月 … 2015年06月</p>
<p>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</p> <p>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】</p>
<p>★★★★★ 4</p> <p>投稿者さんの 鶴川シーワールドホテル のクチコミ (感想・情報)</p> <p> 投稿者さん 2015年06月11日 07:33:49</p> <p>夫、2歳半と5ヶ月の子どもの4人で宿泊しました。</p> <p>【立地】 当たり前ですが鶴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。</p> <p>【部屋】 至って普通です。(古いからか、勝手の声は少し聞こえます。) トイレ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。</p> <p>【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。</p> <p>【風呂】 小さな子ども(赤ちゃん)用のグッズ(ベビーベッド、コーナー、バス、おもちゃ、泡ソーブ、支架のあるイス)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での、温泉を楽しむという雰囲気ではなく、錢湯の如湯が温泉という感じです。</p> <p>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。</p> <p>【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むと快く入</p> <p> 鶴川シーワールドホテル 2015年06月11日 19:25:48</p> <p>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。</p> <p>詳細にご感想頂きまして、ありがとうございました。</p> <p>今後の参考にさせて頂きます。</p> <p>また、スタッフ対応に関して、お褒めのお言葉を頂戴しました、</p> <p>とても嬉しく思います。</p> <p>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、</p> <p>スタッフと共にさせて頂きます。</p> <p>最後に、「アメニティ・シャンプー」の件、大変申し訳ございませんでした。</p> <p>早急に対応をして、改善を行います。</p> <p>貴重なご意見を、ありがとうございます。</p> <p>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。</p>
<p>いい値パリュープラン</p> <p>【最安料金 (日安)】 10,186円~ (消費税込11,000円~)</p> <p>【当日15:50からアシカと記念撮影】笑うアシカと一緒にパリ付プラン 室数限定 【最安料金 (日安)】 10,278円~ (消費税込11,100円~)</p> <p>【当日14:40~エコ・アクアロームコミュニケーションタイム 1日3組限】 【最安料金 (日安)】 10,278円~ (消費税込11,100円~)</p> <p>【夜の水族館探検付】3月~10月の火・木曜日限定プラン 【最安料金 (日安)】 10,278円~ (消費税込11,100円~)</p> <p>【当日14:50からイルカと一緒にパリ付】2室限 【最安料金 (日安)】 10,463円~ (消費税込11,300円~)</p> <p>今しかない!★アワーライ 料付サマーワールド大園 バスケット付で大満足 5月・6月の木~木曜日 限定プラン 【最安料金 (日安)】 10,926円~ (消費税込11,800円~)</p> <p>【便利な赤ちゃんグッズ付】初お泊りはお母さんも嬉しい★赤ちゃんな つ得プラン 【最安料金 (日安)】 10,926円~ (消費税込11,800円~)</p> <p>お子様にも大好評!オ ンシャンプーブラン 【最安料金 (日安)】 11,112円~ (消費税込12,000円~)</p> <p>【80cmのジャンボボイ ズ】海の王者シャチぬいぐるみ付プラン 【最安料金 (日安)】 11,204円~ (消費税込12,100円~)</p> <p>房総2大テーマパーク満 喫「マザーフラウドチケッ ト」付プラン 【最安料金 (日安)】 11,389円~ (消費税込12,300円~)</p> <p>【当日14:50~イルカ</p>

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

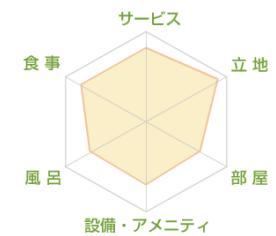
評価内訳

- 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。

テキストデータ

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。

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

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

データ理解

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 2022-2023 および 2024-2025(～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)

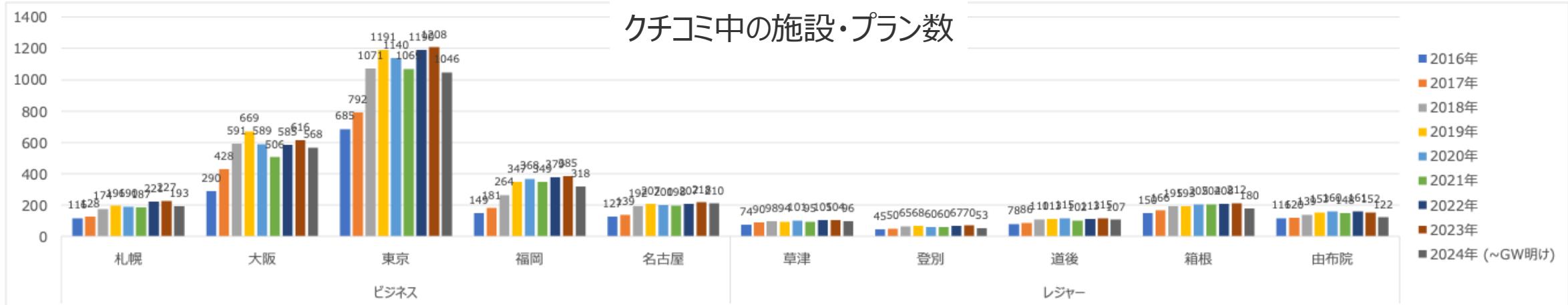
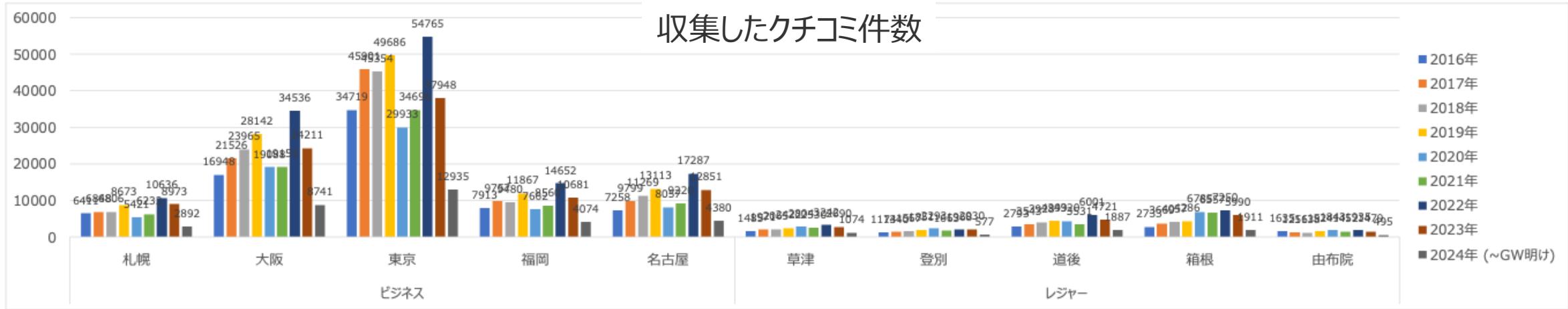
● 楽天トラベル のクチコミデータ

- 収集期間は 2022-2023 および 2024-2025(～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

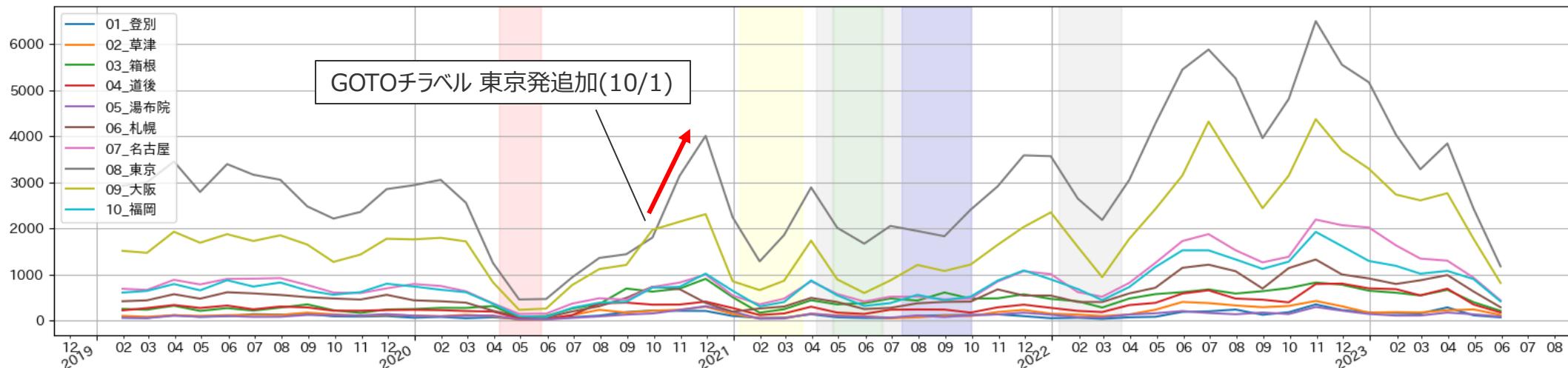
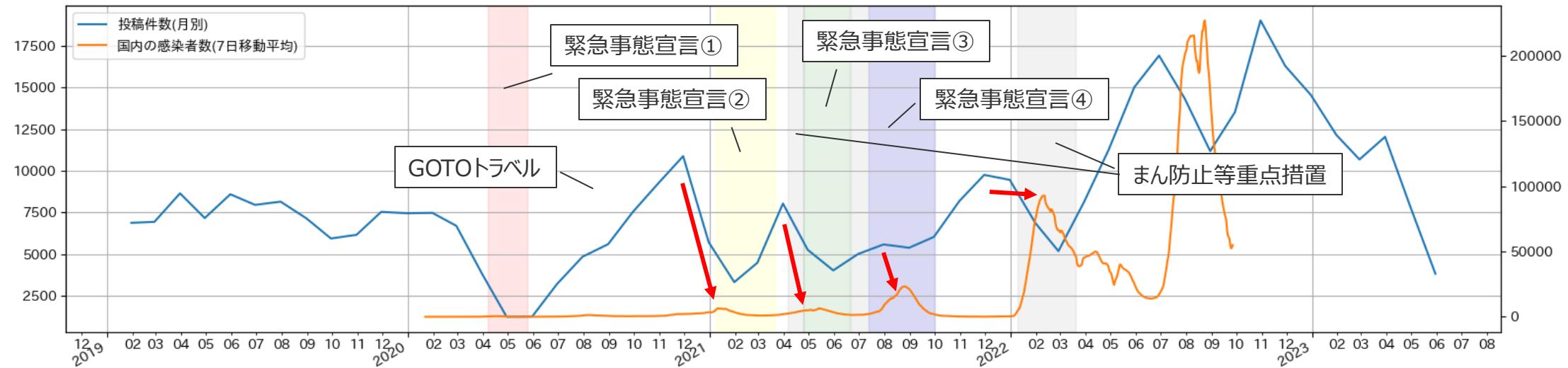
参考 — Webサイトクローリング

- 全量では 172.0万件、2022-2023は21.0万件、2024-2025は13.8万件



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

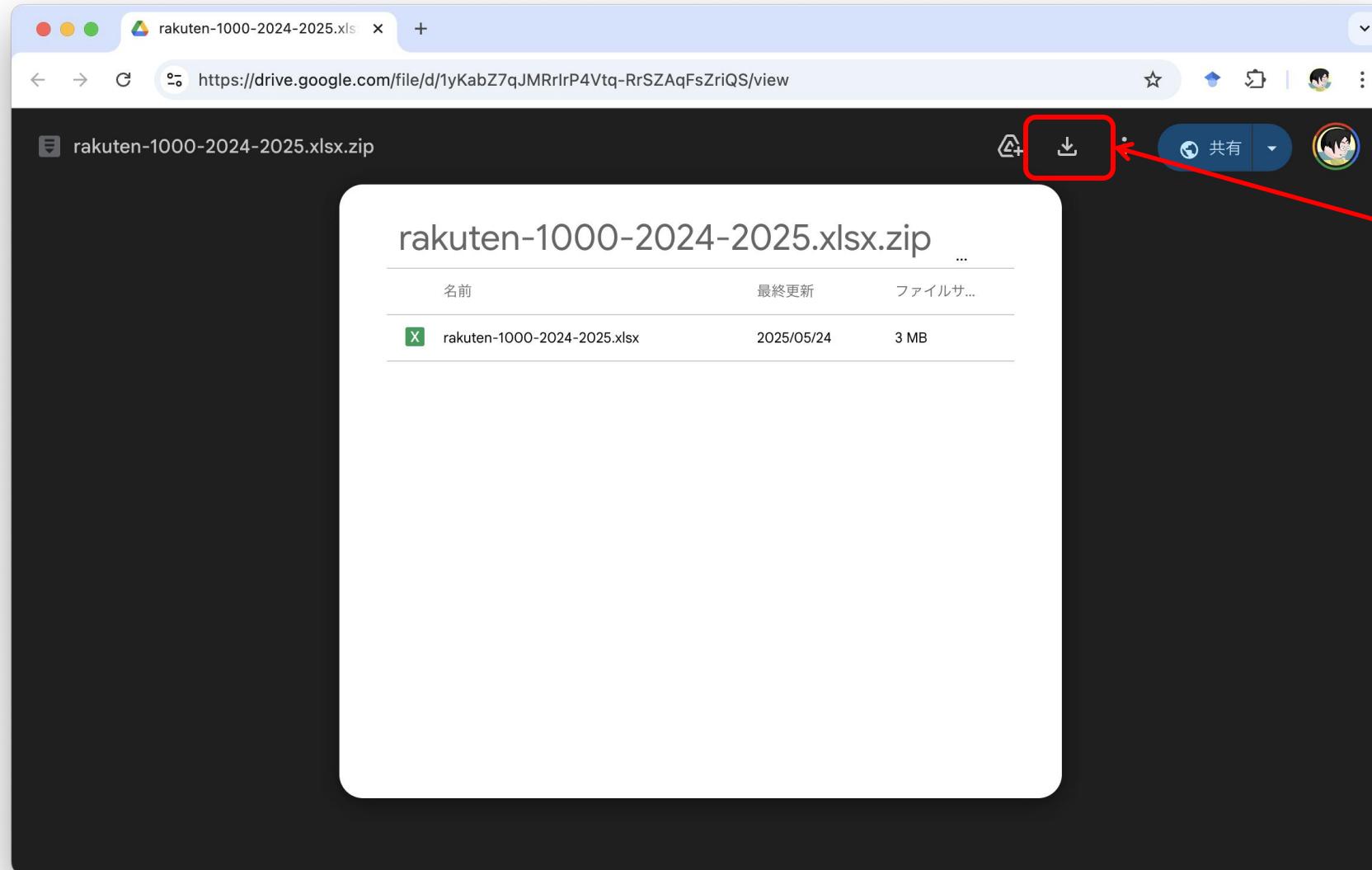


実習用データ — ファイル一覧

● 実習用データは以下の通り → 主に「**rakuten-1000-2024-2025.xlsx**」を使用する

ファイル名	件数 (サイズ)	データセット	備考
<u>rakuten-1000-2024-2025.xlsx.zip</u>	10,000 (2.5 MB)	<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: 2024/1~2025 GW明け	本講義の全体を通して使用する
<u>rakuten-1000-2022-2023.xlsx.zip</u>	10,000 (2.5 MB)	<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: 2022/1~2023/12	演習用 (年度で比較する場合など)
<u>rakuten-all-2024-2025-tsv.zip</u>	138,214 (17.4 MB)	<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: 2024/1~2025 GW明け	参考用
<u>rakuten-all-2022-2023-tsv.zip</u>	209,603 (26.8 MB)	<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: 2022/1~2023/12	参考用
<u>rakuten-all-tsv.zip</u>	1,720,202 (225 MB)	<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: 2009/3~2024 GW明け	参考用

参考 — Google Drive ダウンロード画面



ここをクリックすると
ダウンロードが始ま
ります

テキストマイニングの手順

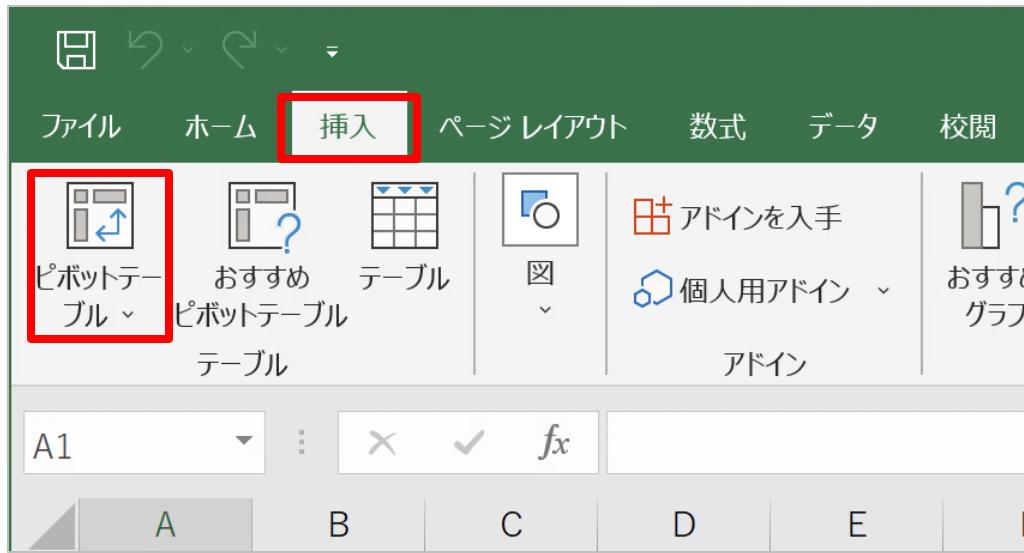
- データをよく知る
 - データ件数や構成比を集計 → データを理解する
 - 旅行目的別の人気エリアは?
 - 同伴者別の人気エリアは?
 - 数値評価による人気エリアの差異は?
- テーマを設定する
 - 解決すべき課題を決める → 分析目的を明確にする
 - 数値評価が低い原因は?
 - 高評価の施設に学ぶ改善点は?
- テキスト分析に取り組む
 - これら課題を解決するために、テキスト分析を実施

データ理解 — EXCEL を使った集計

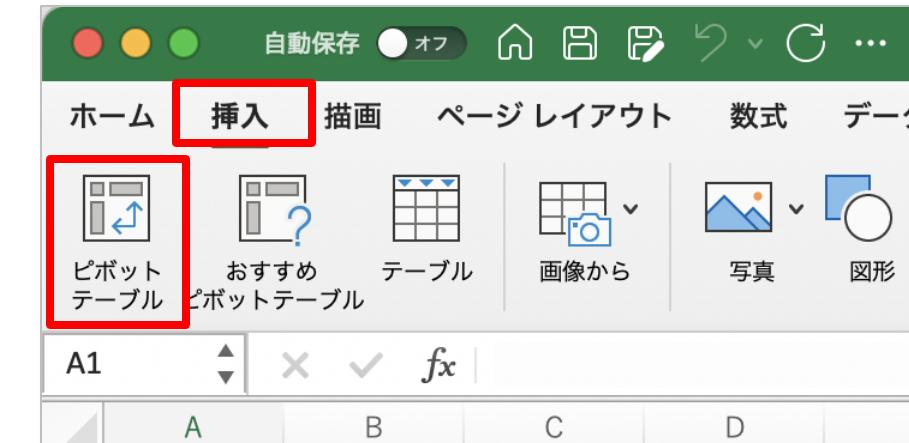
● EXCEL のピボットテーブルを使ってデータを集計する

- ① ファイル **rakuten-1000-2024-2025.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 / コメント
■ 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 / コメント	列ラベル	行ラベル	男性	女性	na	総計
■ A_レジャー	5000	10代		0.01%	0.04%	0.00%	0.05%
01_登別	1000	20代		0.58%	1.14%	0.00%	1.72%
02_草津	1000	30代		2.00%	2.40%	0.00%	4.40%
03_箱根	1000	40代		4.15%	3.22%	0.00%	7.37%
04_道後	1000	50代		8.06%	4.06%	0.00%	12.12%
05_湯布院	1000	60代		6.33%	2.61%	0.00%	8.94%
■ B_ビジネス	5000	70代		1.54%	0.46%	0.00%	2.00%
06_札幌	1000	80代		0.07%	0.04%	0.00%	0.11%
07_名古屋	1000	90代		0.01%	0.00%	0.00%	0.01%
08_東京	1000	120代		0.00%	0.01%	0.00%	0.01%
09_大阪	1000	na		0.00%	0.00%	63.27%	63.27%
10_福岡	1000	総計		22.75%	13.98%	63.27%	100.00%

③投稿者の傾向 (性別xカテゴリ別)

行ラベル	個数 / コメント	列ラベル	行ラベル	A_レジャー	B_ビジネス	総計
■ A_レジャー	5000	■ 男性	男性	21.52%	23.98%	22.75%
■ B_ビジネス	5000	■ 女性	女性	16.80%	11.16%	13.98%
na	na	na	na	61.68%	64.86%	63.27%
総計	10000	総計	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
		A_レジャー					B_ビジネス								
行ラベル	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	24.40%	21.20%	15.80%	26.50%	19.70%	21.52%	25.90%	26.00%	20.80%	23.60%	23.60%	23.98%		22.75%	
女性	16.80%	17.40%	18.30%	13.00%	18.50%	16.80%	10.90%	9.90%	11.60%	10.50%	12.90%	11.16%		13.98%	
na	58.80%	61.40%	65.90%	60.50%	61.80%	61.68%	63.20%	64.10%	67.60%	65.90%	63.50%	64.86%		63.27%	
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%	

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
		A_レジャー					B_ビジネス								
行ラベル	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	0.00%	0.10%	0.20%	0.00%	0.00%	0.06%	0.10%	0.00%	0.00%	0.10%	0.00%	0.04%		0.05%	
20代	1.90%	2.20%	3.50%	0.90%	1.90%	2.08%	0.80%	1.70%	1.40%	1.60%	1.30%	1.36%		1.72%	
30代	4.80%	4.90%	6.00%	3.90%	5.70%	5.06%	4.20%	3.60%	3.40%	3.70%	3.80%	3.74%		4.40%	
40代	7.10%	6.80%	5.80%	7.50%	7.20%	6.88%	8.20%	9.10%	8.30%	7.20%	6.50%	7.86%		7.37%	
50代	15.30%	12.60%	8.10%	12.80%	11.60%	12.08%	12.40%	11.30%	10.70%	13.00%	13.40%	12.16%		12.12%	
60代	9.70%	9.30%	7.10%	11.00%	9.30%	9.28%	9.60%	9.00%	7.50%	7.20%	9.70%	8.60%		8.94%	
70代	2.20%	2.50%	3.10%	3.20%	2.40%	2.68%	1.30%	1.10%	1.10%	1.30%	1.80%	1.32%		2.00%	
80代	0.20%	0.20%	0.20%	0.20%	0.10%	0.18%	0.20%	0.00%	0.00%	0.00%	0.00%	0.04%		0.11%	
120代	0.00%	0.00%	0.10%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%		0.01%	
na	58.80%	61.40%	65.90%	60.50%	61.80%	61.68%	63.20%	64.10%	67.60%	65.90%	63.50%	64.86%		63.27%	
90代	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	0.02%		0.01%	
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	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				
一人		26.70%	18.00%	14.30%	49.70%	15.30%	24.80%	57.60%	65.90%	69.10%	62.80%	55.60%	62.20%		43.50%
家族		60.80%	63.50%	64.50%	38.30%	64.70%	58.36%	31.70%	25.30%	21.00%	26.20%	32.50%	27.34%		42.85%
恋人		5.50%	10.50%	12.00%	3.90%	9.90%	8.36%	2.90%	3.30%	3.60%	3.10%	3.10%	3.20%		5.78%
友達		4.70%	6.50%	7.80%	3.50%	8.90%	6.28%	5.10%	3.40%	4.60%	6.20%	5.30%	4.92%		5.60%
仕事仲間		1.70%	0.70%	0.50%	3.70%	0.50%	1.42%	2.00%	1.40%	1.10%	1.40%	3.10%	1.80%		1.61%
その他		0.60%	0.80%	0.90%	0.90%	0.70%	0.78%	0.70%	0.70%	0.60%	0.30%	0.40%	0.54%		0.66%
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%

⑦数値評価の構成 (総合別xカテゴリ-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				
5		43.10%	47.80%	52.30%	47.70%	67.30%	51.64%	43.70%	37.70%	39.70%	42.70%	37.90%	40.34%		45.99%
4		39.60%	38.40%	34.90%	38.90%	24.20%	35.20%	40.00%	43.60%	42.80%	41.30%	43.30%	42.20%		38.70%
3		10.50%	8.00%	7.30%	9.80%	5.00%	8.12%	11.60%	13.60%	12.10%	11.60%	13.40%	12.46%		10.29%
2		4.30%	3.80%	3.80%	2.60%	2.50%	3.40%	3.50%	2.90%	3.40%	2.40%	3.90%	3.22%		3.31%
1		2.50%	2.00%	1.70%	1.00%	1.00%	1.64%	1.20%	2.20%	2.00%	2.00%	1.50%	1.78%		1.71%
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

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	4.26	4.28	4.16	4.06	4.28	4.30	4.32
01_登別	4.08	4.24	3.98	3.94	4.28	4.13	4.17
02_草津	4.20	4.27	4.06	3.94	4.30	4.20	4.26
03_箱根	4.30	4.16	4.21	4.06	4.28	4.37	4.32
04_道後	4.20	4.35	4.10	4.02	4.08	4.20	4.30
05_湯布院	4.54	4.38	4.44	4.33	4.47	4.58	4.54
■B_ビジネス	4.01	4.33	4.04	3.90	3.74	4.06	4.16
06_札幌	4.11	4.36	4.11	3.96	3.81	4.13	4.22
07_名古屋	3.98	4.26	3.99	3.88	3.76	3.97	4.12
08_東京	4.00	4.40	4.01	3.92	3.67	4.10	4.15
09_大阪	4.01	4.35	4.08	3.90	3.79	4.02	4.20
10_福岡	3.94	4.28	4.01	3.85	3.69	4.08	4.12

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	4.26	4.28	4.16	4.06	4.28	4.30	4.32
B_ビジネス	4.01	4.33	4.04	3.90	3.74	4.06	4.16

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	4.41	4.43	4.34	4.29	4.44	4.43	4.50
男性	4.33	4.37	4.25	4.15	4.35	4.37	4.47
女性	4.46	4.46	4.40	4.37	4.49	4.46	4.51
■ B_ビジネス	4.15	4.40	4.16	4.02	3.81	4.19	4.21
男性	4.06	4.40	4.07	3.98	3.77	4.18	4.15
女性	4.24	4.39	4.24	4.05	3.85	4.20	4.27

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	4.32	4.38	4.21	4.12	4.34	4.34	4.40
男性	4.20	4.32	4.09	4.02	4.24	4.29	4.33
女性	4.48	4.45	4.36	4.24	4.47	4.41	4.48
■ B_ビジネス	4.03	4.34	4.07	3.92	3.80	4.08	4.22
男性	3.95	4.28	4.00	3.85	3.72	3.97	4.16
女性	4.22	4.48	4.26	4.08	3.97	4.32	4.37

⑨-c 数値評価の平均 (60~80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	4.21	4.23	4.12	4.00	4.27	4.23	4.25
男性	4.18	4.22	4.11	3.97	4.25	4.22	4.25
女性	4.27	4.24	4.15	4.08	4.30	4.24	4.27
■ B_ビジネス	3.91	4.33	4.01	3.76	3.76	3.93	4.13
男性	3.81	4.30	3.90	3.70	3.70	3.89	4.08
女性	4.21	4.41	4.38	3.96	3.92	4.08	4.31

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">・・・	

- 個人ワーク (~19:50 ※休憩込み)
 - 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、検討する
 - 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

(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	4.26	4.28	4.16	4.06	4.20	4.30	4.32	
01_登別	4.08	4.24	3.98	3.94	4.28	4.13	4.17	
02_草津	4.20	4.27	4.06	3.94	4.30	4.20	4.26	
03_箱根	4.30	4.16	4.21	4.06	4.28	4.37	4.32	
04_道後	4.20	4.35	4.10	4.02	4.08	4.20	4.30	
05_湯布院	4.54	4.38	4.44	4.00	4.00	4.00	4.54	
B_ビジネス	4.01	4.33	4.04	4.00	4.16	4.22	4.12	
06_札幌	4.11	4.36	4.11	4.00	4.10	4.15	4.20	
07_名古屋	3.98	4.26	3.99	3.88	3.76	4.00	4.12	
08_東京	4.00	4.40	4.01	3.92	3.67	4.10	4.15	
09_大阪	4.01	4.35	4.08	3.90	3.79	4.02	4.20	
10_福岡	3.94	4.28	4.01	3.85	3.99	4.08	4.12	

- 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	4.26	4.28	4.16	4.06	4.28	4.30	4.32
B_ビジネス	4.01	4.33	4.04	3.90	3.74	4.06	4.16

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	4.26	4.28	4.16	4.06	4.28	4.30	4.32
B_ビジネス	4.01	4.33	4.04	3.90	3.74	4.06	4.16

● 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

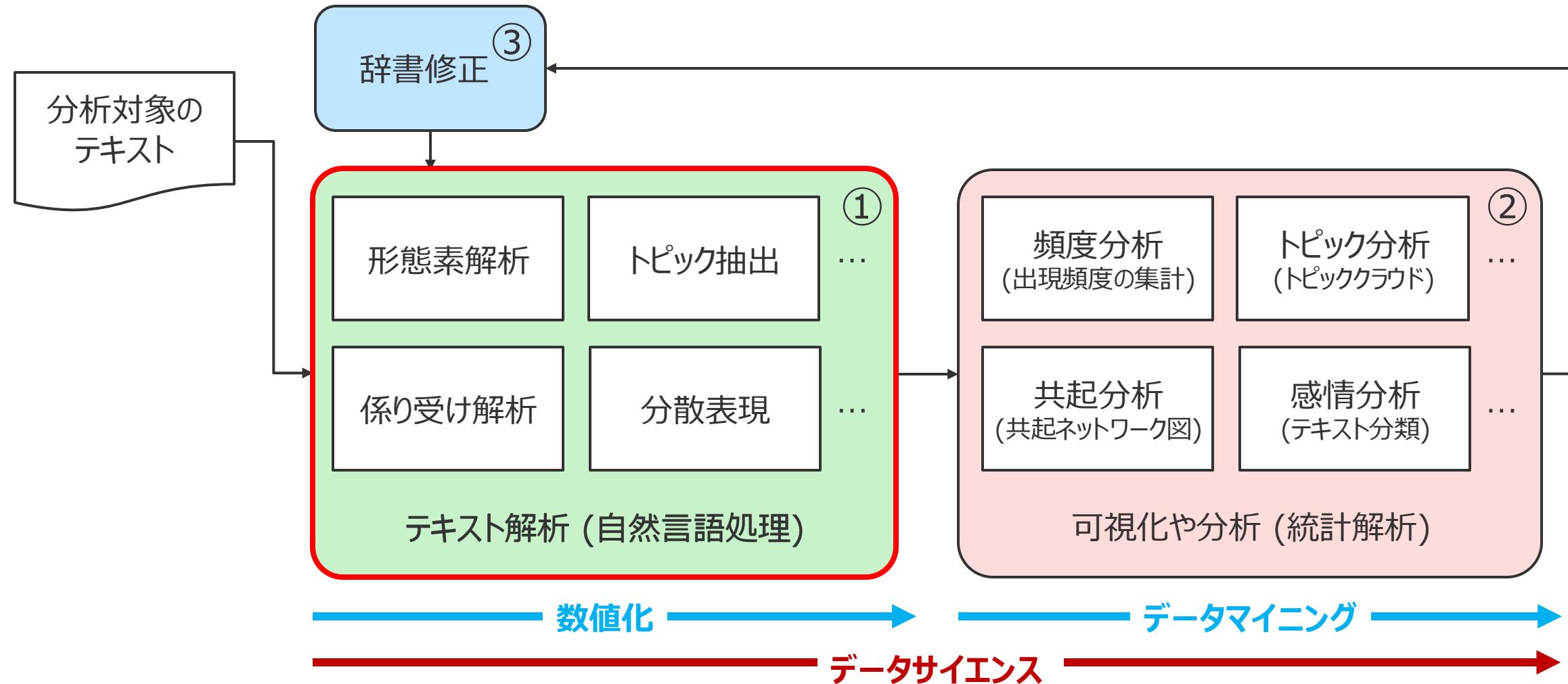
テキスト解析 (1)

(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

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)



伝統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++ が有名

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

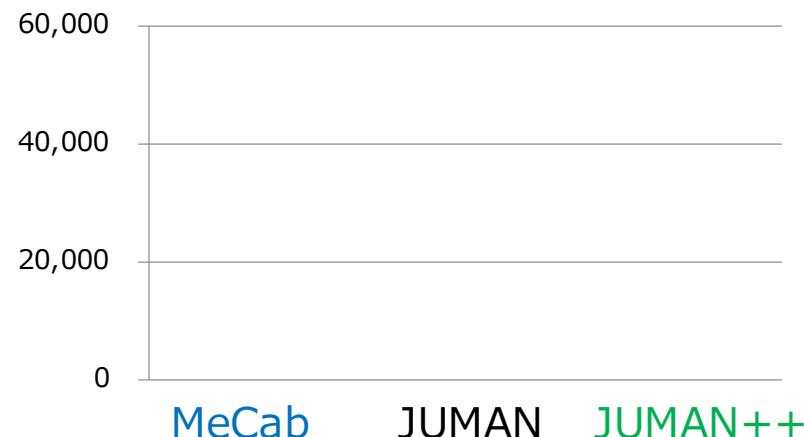
形態素解析器	ChaSen	MeCab	JUMAN	JUMAN++
コスト推定	HMM	CRF	人手	RNNLM
探索方法			接続コスト最小法 (ビタビアルゴリズム)	
係り受け解析	Cabocha	CaboCha		KNP

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/1DVnrsWw4skRgC8jU6_RkeofOQEHFwctc/vie
w?usp=sharing](https://drive.google.com/file/d/1DVnrsWw4skRgC8jU6_RkeofOQEHFwctc/view?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)

次の演習環境 → 注: Googleのアカウントが必要です

● 機械学習の教育・研究を目的とした研究用ツール

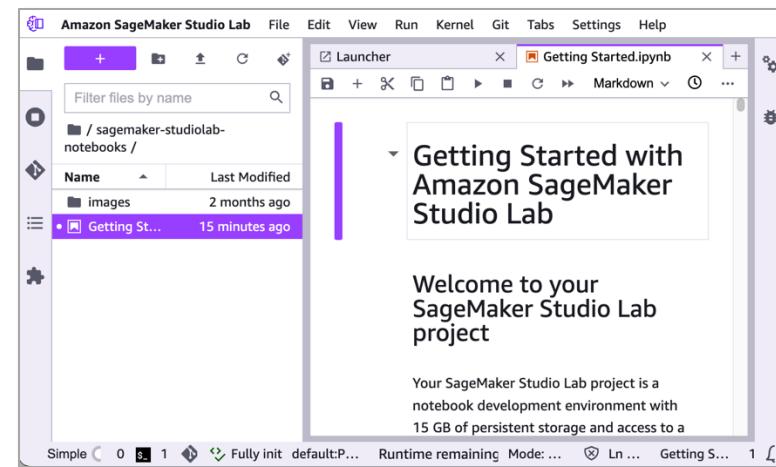
 Colaboratory

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



 Amazon SageMaker Studio Lab

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

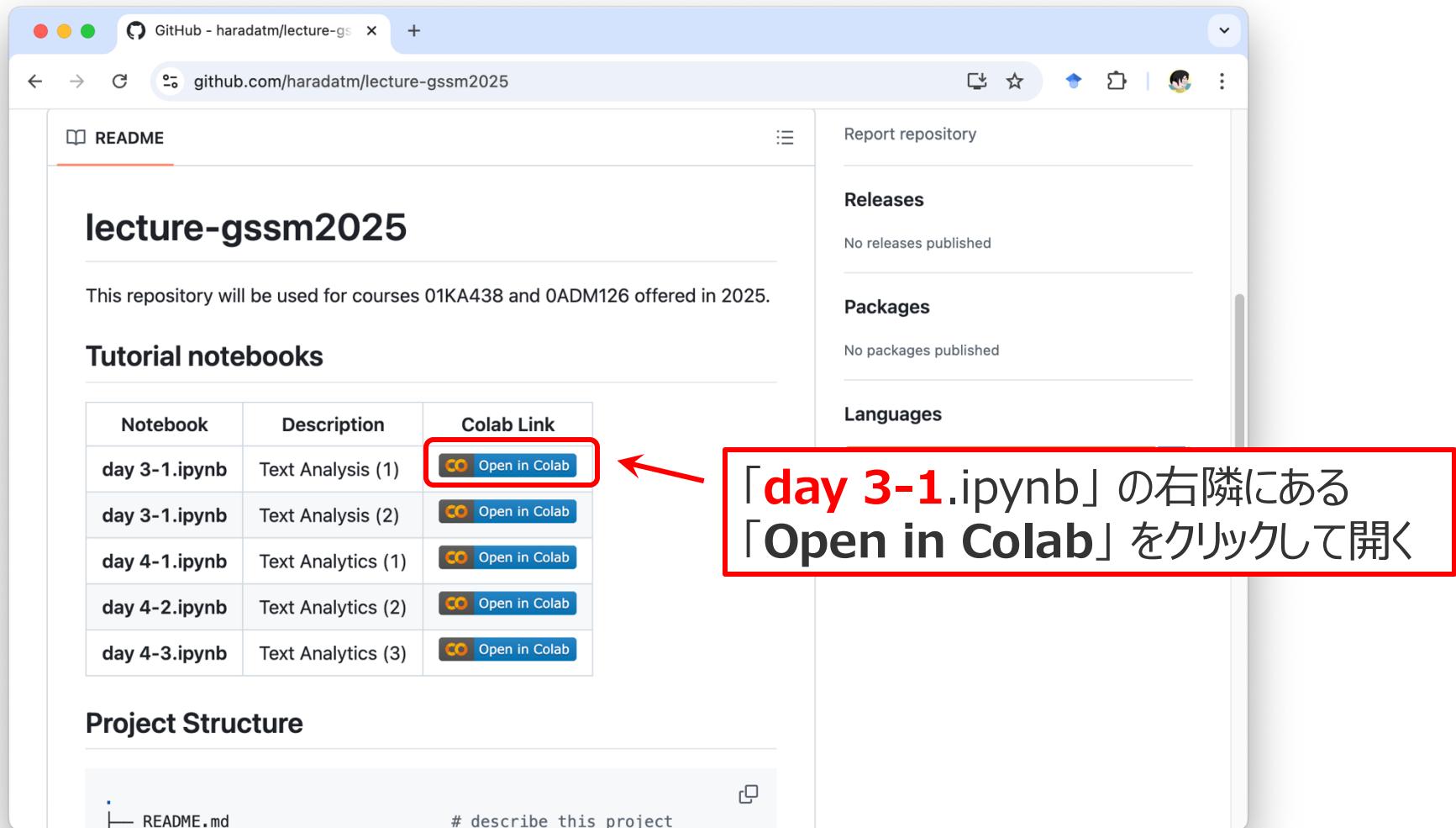


演習で使用
↓

	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)

演習用の教材 – Google Colab ノートブック

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



● 準備

- 0.1 必要なパッケージのインストール
- 0.2 MeCab インストール (時間目安: 約3分)
- 0.3 CaboCha インストール (時間目安: 約4分)
- 0.4 セッションの再起動
- 0.5 動作確認

● テキスト解析 (1)

- 1.1 形態素解析
- 1.2 係り受け解析
- 1.3 辞書追加

目次を開く

目次
day 3-1
準備
0.1 必要なパッケージのインストール
0.2 MeCab インストール (時間目安: 約3分)
0.3 CaboCha インストール (時間目安: 約4分)
0.4 セッションの再起動
0.5 動作確認
テキスト解析 (1)
1.1 形態素解析
1.2 係り受け解析
1.3 辞書追加

● 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

● 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、再生成ボタンを押す

- 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

● 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

● 係り受け解析を行う（係り受けペアを抽出する）

② [セル] # 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 += tree.surface[i]
 return surface

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

← 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

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

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.XX~XX) ※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.XX) ※ページ番号は各スライド右下に記載
- ※ 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