

テキストマイニング

— Part 2 —

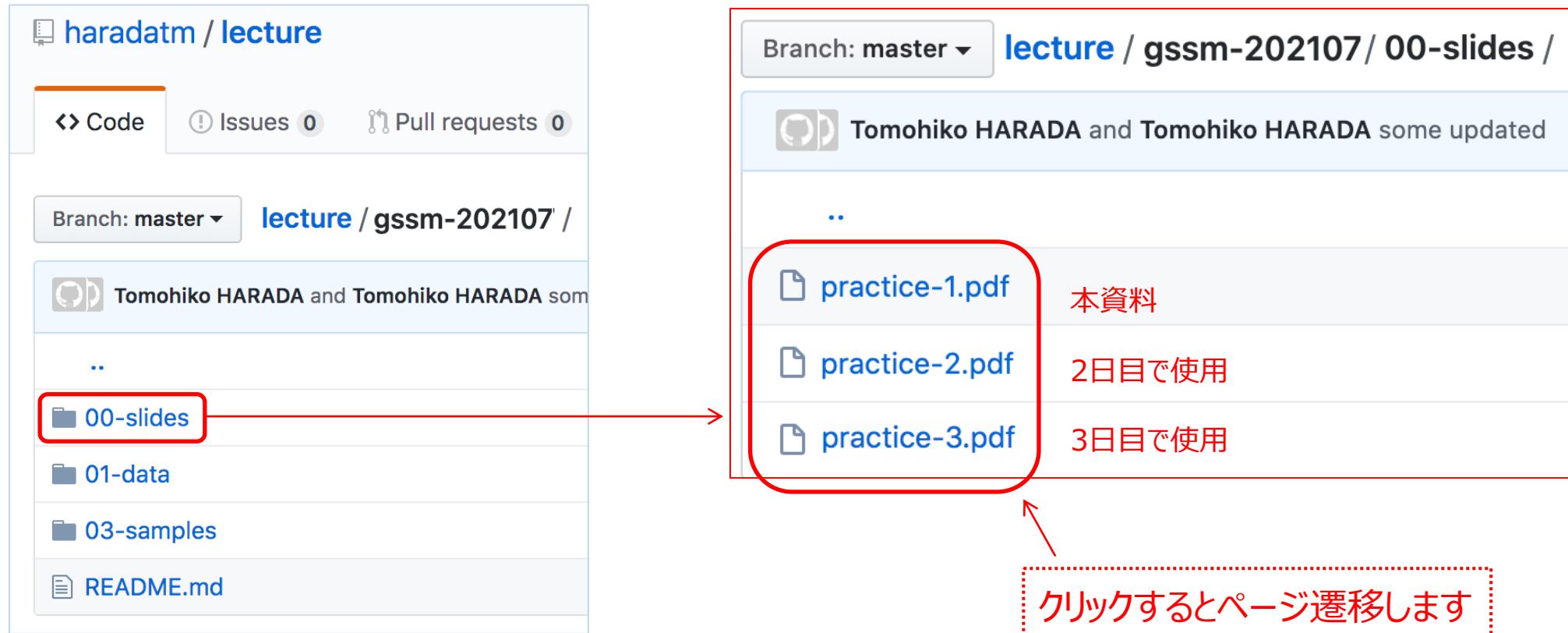
R04年度
人文社会ビジネス科学学術院
ビジネス科学研究群

スケジュール

- Part 1
 - 説明 — 自然言語処理のトレンド
 - 説明 — 環境説明
- Part 2
 - 説明 — テキストマイニングの手順
 - 説明 — データ理解
 - 実習 — データ理解 (Excel)
- Part 3
 - 説明 — データ分析の実践 (KHCoder)
 - 実習 — データ分析の実践 (KHCoder)
- Part 4
 - 実習 — データ分析の実践 (KHCoder)
 - 実習 — データ分析の実践 (KHCoder)

講義スライド

- <https://github.com/haradatm/lecture/tree/master/gssm-202107>



テキストマイニング

- ・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が多い
 - ・震災, 選挙, 新型コロナウィルスなど

事例 – 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



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

事例 — 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
- ・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る
- ・都市型ホテルがどうか変われるか →都市観光ホテル



星野リゾート
ホームページより
(URL はスライド下)



<https://travel.watch.impress.co.jp/docs/news/1056715.html>

<https://www.hoshinoresorts.com/brand/omo/>

口コミ分析の事例をあげてください

クチコミサイトの例



- ホテルのクチコミ数: 1,237万件 ※年間約60~70万

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 a count of 'ホテルのクチコミ数No.1 12,369,840'. Below this, there's a search bar for reviews and a list of new reviews. One review is highlighted with a star rating of 4.03 and another of 4.46.

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 (今回)**
※ 2021/6/4現在

鴨川シーワールドホテル クラ

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

HARADA Tomohiko

楽天宿・航空券・ツアー予約

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 頒賞広場 観光案内

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

鴨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入力に追加 メルマガ 幹事さん機能 友達にメール シェアする 3

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・ディユース 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： 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] 鴨川シーワールドホテル 真っ白なまき もっと見る

施設間連情報

鴨川シーワールドホテル ★トップページ★ 鴨川シーワールドホテル ★鴨シーニュース★

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

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

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

鴨川シーワールドホテルのクチコミ・お客さまの声

総合評価 ★★★★★ 4.12 アンケート件数: 886件

評価内訳 5点 236件 4点 302件 3点 47件 2点 15件 1点 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

サービス立地部屋設備・アメニティ風呂食事

投稿の種類 クチコミ（感想・情報） クチコミ（苦情） 年代 性別 指定なし 男性 挿り込みを解除

宿泊プラン一覧

【1泊朝食付＆カモシー入園バス付】朝からカモシヘルテ's Go! 【最安料金（目安）】 7,963円～ (消費税込8,600円～)

【1日バスポート/朝食付+翌日のランチ券付】遅く着いてもたっぷり楽しめる！ 【最安料金（目安）】 8,519円～ (消費税込9,200円～)

【1泊朝食付】カモシ入園なし ビジネスプラン 【最安料金（目安）】 9,352円～ (消費税込10,100円～)

【カモシオリジナルポンチョ付】6月のお得Days★雨だってへっちゃら♪ 【最安料金（目安）】 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ボーナスプラン1日8室限定 【最安料金（目安）】 9,815円～ (消費税込10,600円～)

【30日前までの予約】早めに決めてお得♪プラン 【最安料金（目安）】 10,000円～

クチコミを投稿する クチコミを修正する

旅行の目的 … レジャー 同伴者 … 家族 審査年月 2015年06月

ご利用の宿泊 【洋室 禁煙・特別室】お部屋からシャチャイルカも見える 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】】

総合 ★★★★★ 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 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、それはいまいせでした。
気にかかる事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
です。

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。

鴨川シーワールドホテル クラ

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

HARADA Tomohiko

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

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した。
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（目安）】 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パリ付プラン 室数限定

【最安料金（目安）】 10,278円～ (消費税込11,100円～)

【当日13:40～エコ・アクアロームコミュニケーションタイム 1日3組限定】

【最安料金（目安）】 10,278円～ (消費税込11,100円～)

【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（目安）】 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付 2室限定】鴨川シーワールド休憩付プラン

【最安料金（目安）】 10,463円～ (消費税込11,300円～)

今しかない！★アワビ料理付&シーワールド入園バスポート付で大満足♪5月・6月の月～木曜日限定プラン

【最安料金（目安）】 10,926円～ (消費税込11,800円～)

【便利な赤ちゃんグッズ付】初！お泊りはお母さんも嬉しい★赤ちゃんが育つプラン

【最安料金（目安）】 10,926円～ (消費税込11,800円～)

お子様にも大好評！オーシャンピューブラン

【最安料金（目安）】 11,112円～ (消費税込12,000円～)

【80cmのジャッポサイズ】海の王者シャチぬいぐるみ付プラン

【最安料金（目安）】 11,204円～ (消費税込12,100円～)

房費2大テマパーク満喫「マザーパークチェック」付プラン

【最安料金（目安）】 11,389円～ (消費税込12,300円～)

9

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 19:32:50

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した。

旅館の件、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した。
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。

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

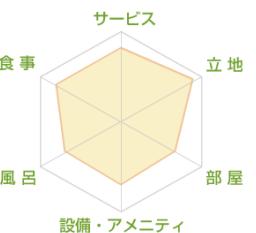
評価内訳

- 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）



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

情報



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?

・データ分析に取り組む

- ・これら課題を解決するために、テキスト分析を実施

実習で使用するデータ

楽天トラベル のクチコミデータ

- ・収集期間は **2019-2020** および **2021-2022(～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 = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

実習で使用するデータ

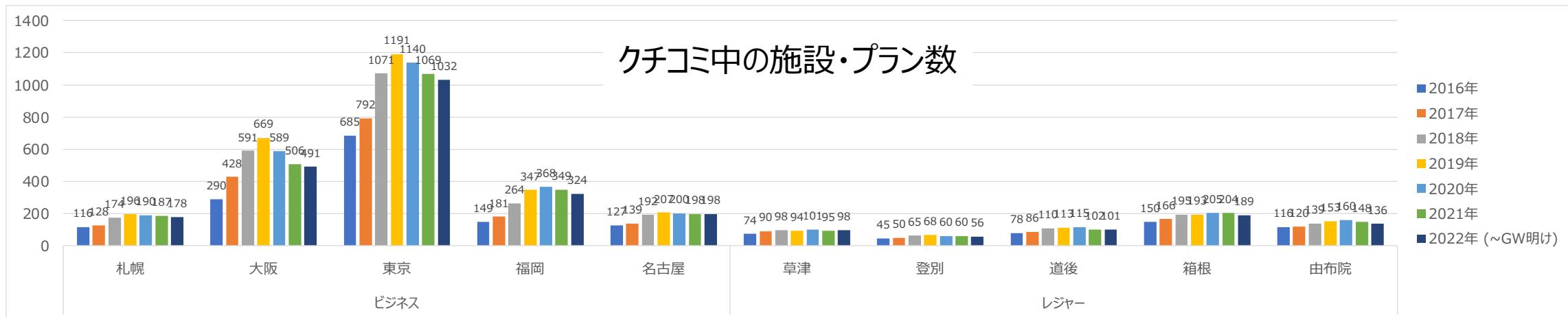
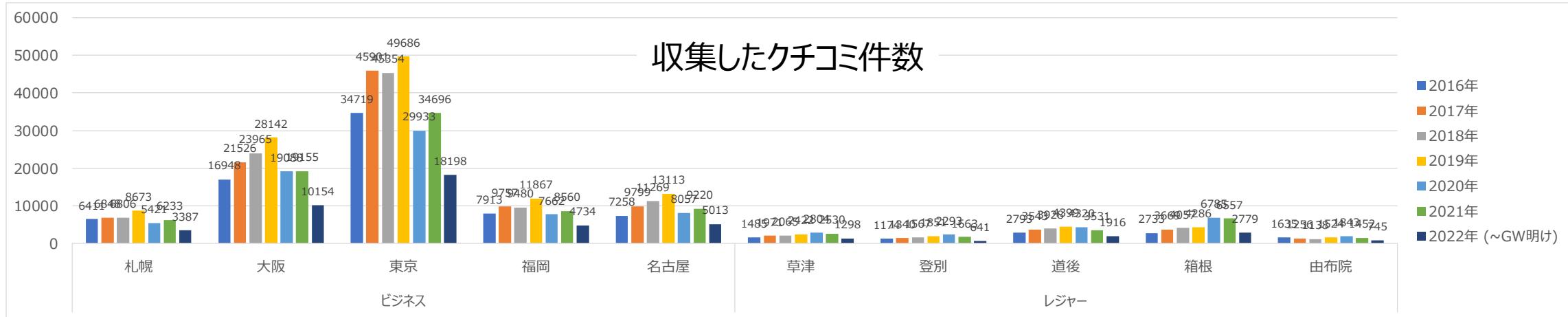
楽天トラベル のクチコミデータ

- 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**)

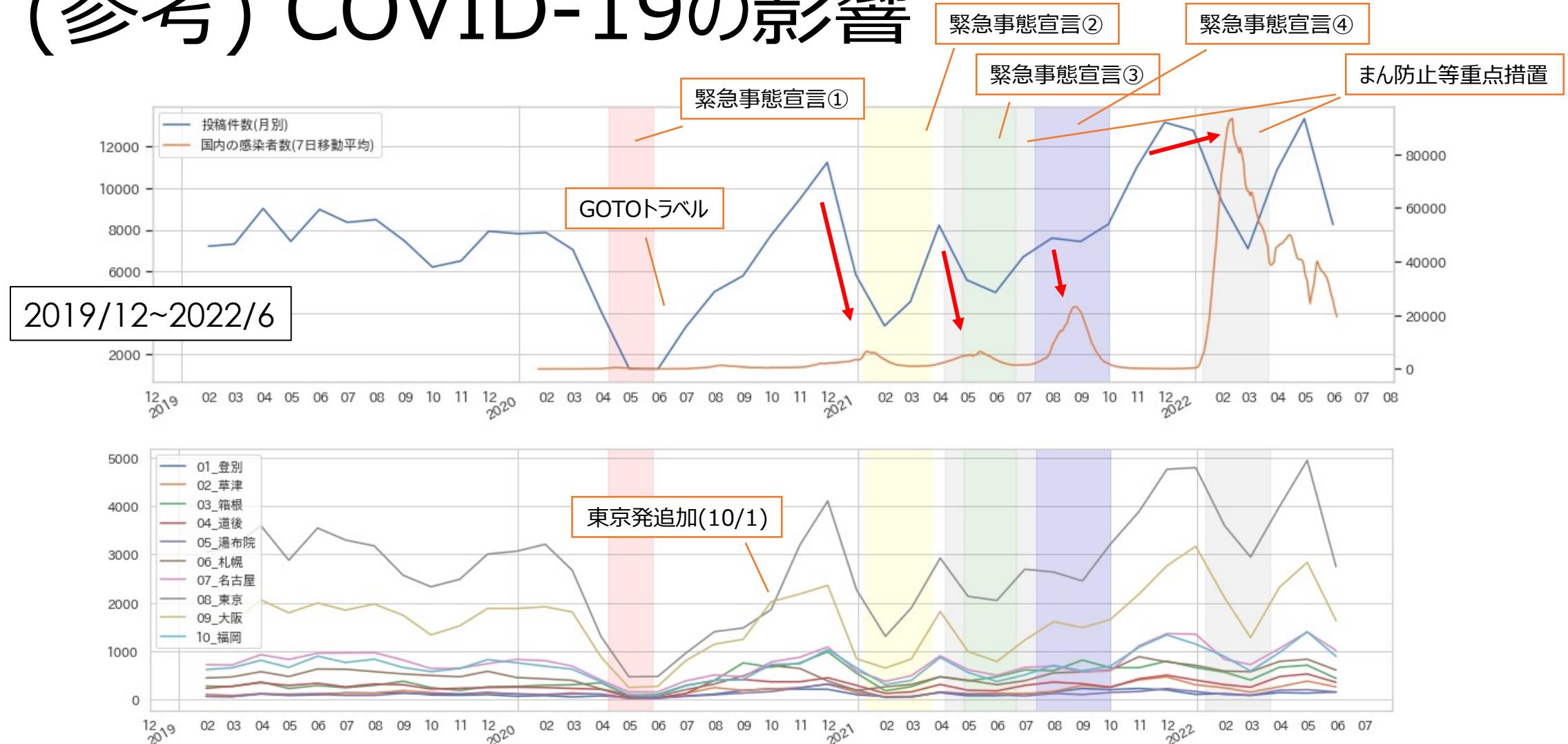
施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント (テキスト)
ユーザー評価	7項目	総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目	旅行の目的, 同伴者
宿泊日	1項目	宿泊年月
ユーザー情報	3項目	ユーザー, 年代, 性別

(参考) 全収集データの推移

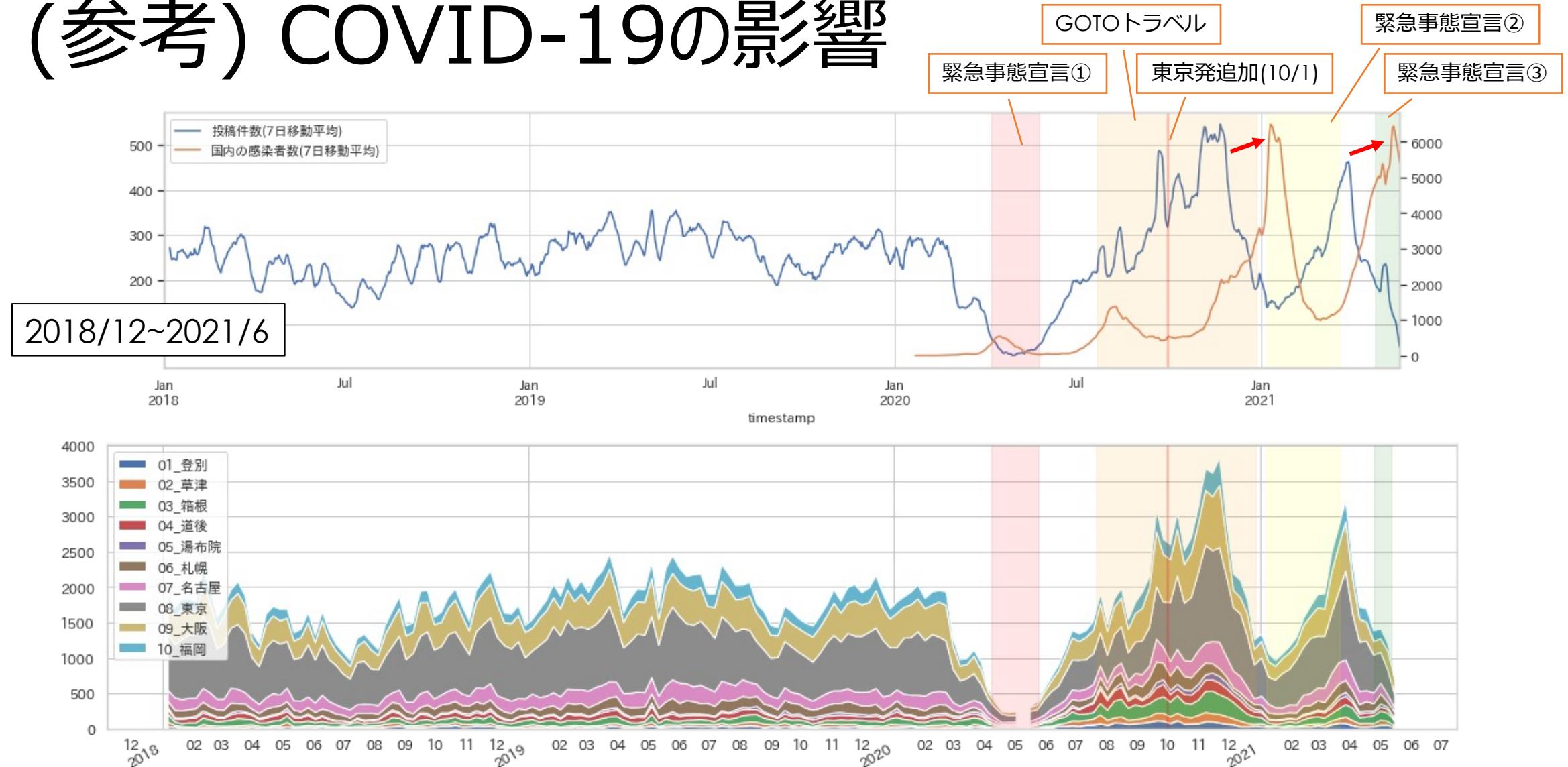
※ 実習では 2020~2022年のエリアごとに
サンプリングした1,000件を使用



(参考) COVID-19の影響



(参考) COVID-19の影響

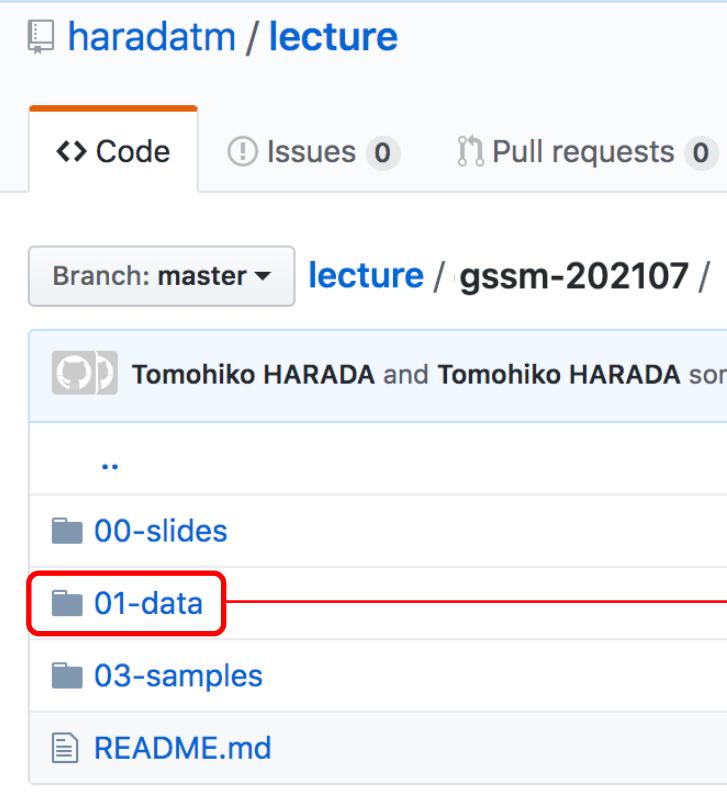


データ一覧

データファイル名	件数	データセット	備考
rakuten-1000-2021-2022.xlsx	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: 2020/1~2022/5/15	<ul style="list-style-type: none">本講義の全体を通して利用する
rakuten-1000-2019-2020.xlsx	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: 2019/1~2020/12	<ul style="list-style-type: none">実習用 (期間で比較する等)
rakuten-all-2021-2022-tsv.zip	142,475	<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 (宿泊年月naを除く)期間: 2020/1~2022/5/15	<ul style="list-style-type: none">その他 (Python や R を使って分析したい人向け)
rakuten-all-2019-2020-tsv.zip	162,433	<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 (宿泊年月naを除く)期間: 2019/1~2020/12	
rakuten-all-tsv.zip	1,593,525	<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	

データの取得方法

- <https://github.com/haradatm/lecture/tree/master/gssm-202207>



haradatm / lecture

Code Issues 0 Pull requests 0

Branch: master lecture / gssm-202107 /

Tomohiko HARADA and Tomohiko HARADA som

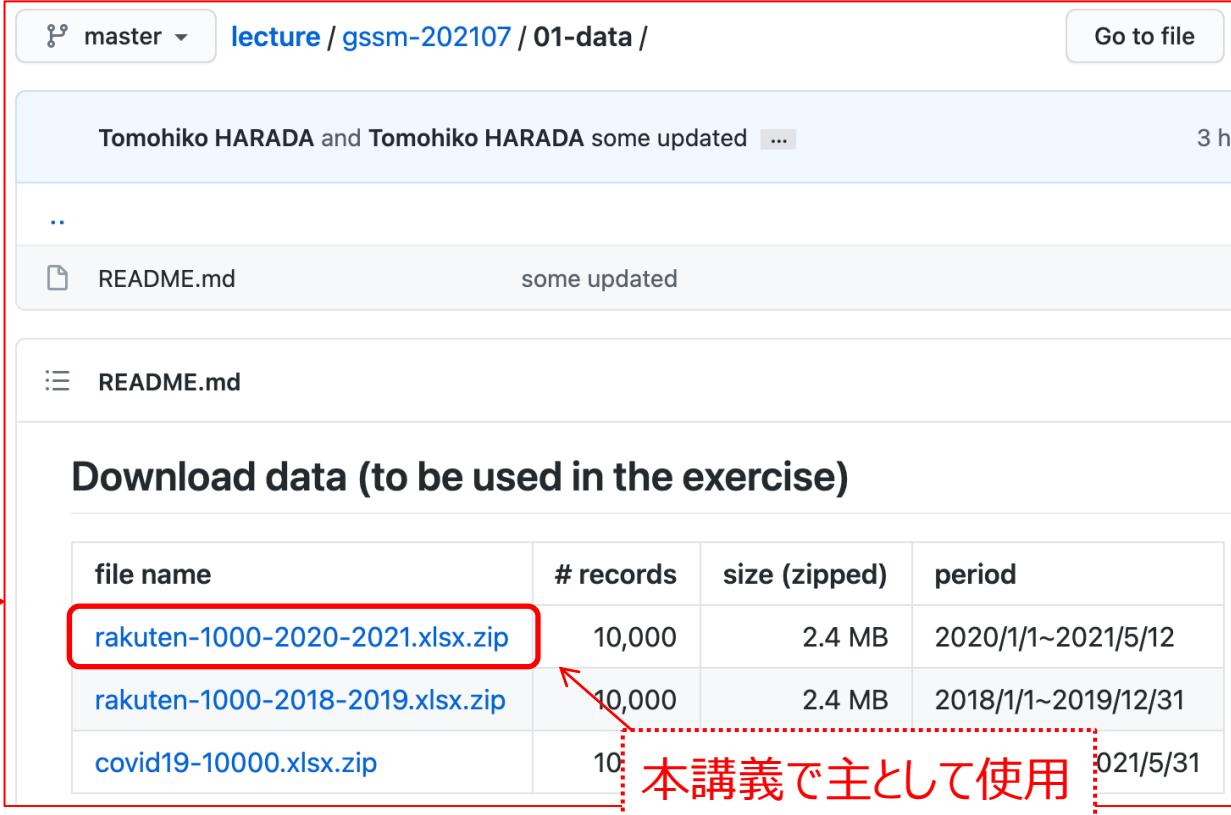
..

00-slides

01-data

03-samples

README.md



master lecture / gssm-202107 / 01-data / Go to file

Tomohiko HARADA and Tomohiko HARADA some updated ... 3 hc

..

README.md some updated

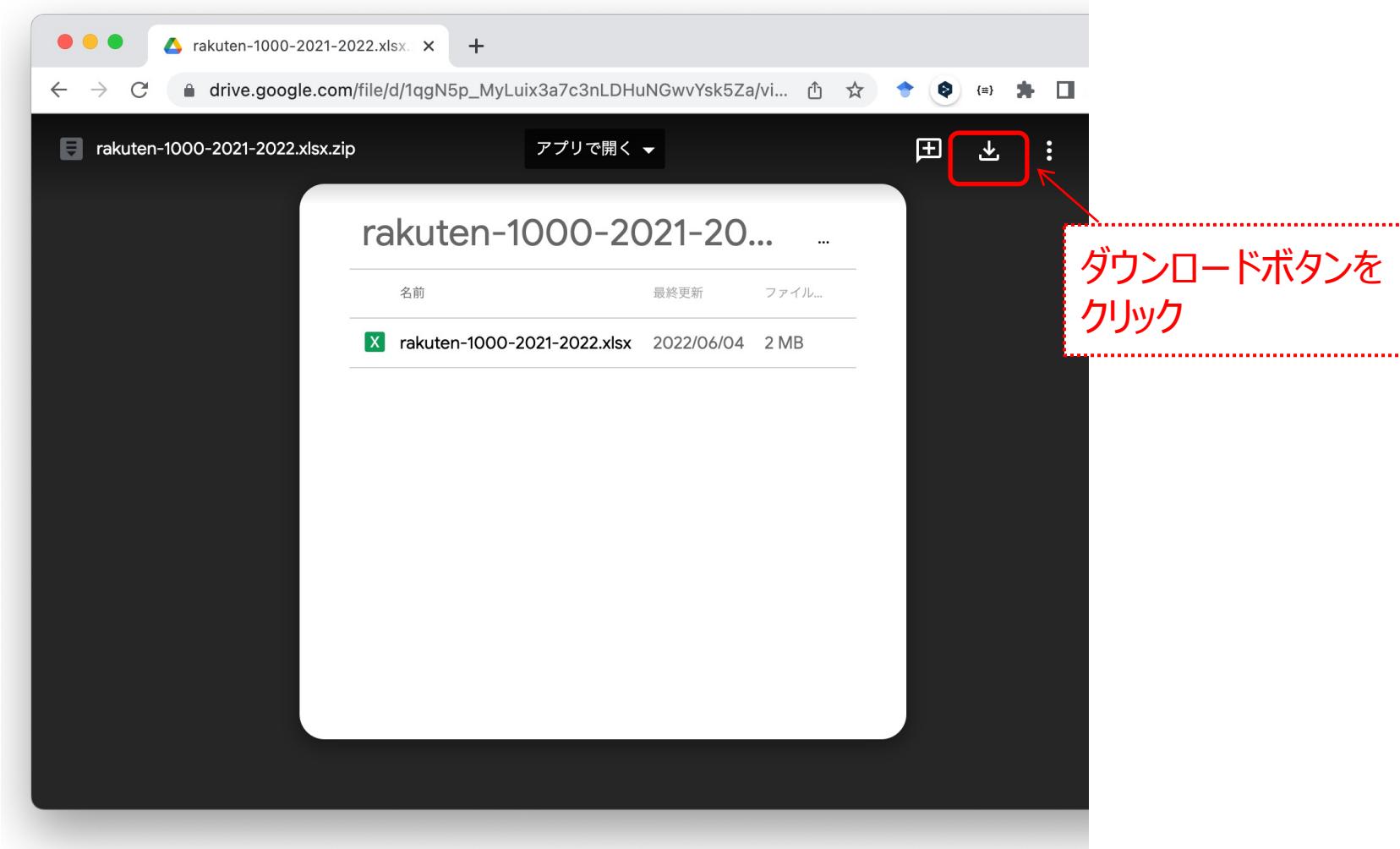
README.md

Download data (to be used in the exercise)

file name	# records	size (zipped)	period
rakuten-1000-2020-2021.xlsx.zip	10,000	2.4 MB	2020/1/1~2021/5/12
rakuten-1000-2018-2019.xlsx.zip	10,000	2.4 MB	2018/1/1~2019/12/31
covid19-10000.xlsx.zip	10		2021/5/31

本講義で主として使用

ダウンロード方法



データ理解

- ・ピボットテーブル(EXCEL)を使ってデータを集計する
 - ・ファイル **rakuten-1000-2020-2021.xlsx** を開く
 - ・A～R 列を選択し,ピボットテーブルを作成する

【Windows】Excel 2007・2010・2013



[挿入] タブ [テーブル] グループの [ピボットテーブル] ボタンをクリックします

演習 — データ理解

- EXCELを使ってデータ集計を行い,発見した特徴や傾向をもとにデータセットを説明(要約)する

例) データセットを説明する観点

- 投稿者の属性(年代,性別)は?
- 旅行目的別の人気エリアは?
- 同伴者別の人気エリアは?

データ理解 — 集計例

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

投稿者の傾向 (年代別・性別)

行ラベル	個数 / コメン	列ラベル	男性	女性	.	総計
10代			0.03%	0.03%	0.00%	0.06%
20代			1.04%	1.21%	0.00%	2.25%
30代			2.05%	2.14%	0.00%	4.19%
40代			5.66%	3.57%	0.00%	9.23%
50代			9.08%	4.31%	0.00%	13.39%
60代			4.58%	1.53%	0.00%	6.11%
70代			0.92%	0.21%	0.00%	1.13%
80代			0.04%	0.00%	0.00%	0.04%
90代			0.01%	0.00%	0.00%	0.01%
110代			0.00%	0.02%	0.00%	0.02%
120代			0.01%	0.00%	0.00%	0.01%
総計			23.42%	13.02%	63.56%	100.00%

投稿者の傾向 (エリア別)

行ラベル	個数 / コメン	列ラベル
男性	23.36%	23.48%
女性	15.62%	10.42%
.	61.02%	66.10%
総計	100.00%	100.00%
		100.00%

データ理解 — 集計例

投稿者の傾向 (性別, 目的-エリア別)

個数 / コメント	列ラベル	A_レジャー										B_ビジネス					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	A_レジャー 集計					06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
男性		19.40%	22.40%	17.00%	23.70%	19.60%	20.42%	27.10%	29.40%	24.70%	25.80%	26.10%	26.62%	23.52%				
女性		16.60%	15.00%	15.80%	13.50%	18.70%	15.92%	10.40%	7.20%	8.80%	9.50%	9.30%	9.04%	12.48%				
na		64.00%	62.60%	67.20%	62.80%	61.70%	63.66%	62.50%	63.40%	66.50%	64.70%	64.60%	64.34%	64.00%				
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%	100.00%	100.00%	

投稿者の傾向 (年代別, 目的-エリア別)

個数 / コメント	列ラベル	A_レジャー										B_ビジネス					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	A_レジャー 集計					06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
10代		0.00%	0.00%	0.40%	0.10%	0.00%	0.10%	0.00%	0.00%	0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	0.02%	0.06%	
20代		0.80%	3.10%	4.50%	2.10%	3.50%	2.80%	1.90%	1.30%	2.50%	1.70%	1.10%	1.70%	1.10%	1.70%	1.70%	2.25%	
30代		4.80%	5.20%	4.20%	4.30%	6.50%	5.00%	3.50%	3.90%	3.10%	3.40%	3.00%	3.38%	3.00%	3.38%	3.38%	4.19%	
40代		10.30%	9.40%	6.90%	7.80%	9.10%	8.70%	9.60%	11.40%	8.50%	9.00%	10.30%	9.76%	9.76%	9.76%	9.76%	9.23%	
50代		13.20%	13.50%	9.00%	16.40%	14.10%	13.24%	14.20%	12.80%	12.00%	14.00%	14.70%	13.54%	13.54%	13.54%	13.54%	13.39%	
60代		7.60%	6.40%	6.80%	7.50%	9.40%	7.54%	4.80%	3.90%	4.30%	5.70%	4.70%	4.68%	4.68%	4.68%	4.68%	6.11%	
70代		1.30%	1.90%	1.70%	0.90%	1.70%	1.50%	0.80%	1.40%	0.20%	0.70%	0.70%	0.76%	0.76%	0.76%	0.76%	1.13%	
80代		0.00%	0.10%	0.10%	0.10%	0.00%	0.06%	0.00%	0.00%	0.10%	0.00%	0.00%	0.02%	0.02%	0.02%	0.02%	0.04%	
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%	0.02%	0.02%	0.02%	0.01%	
110代		0.00%	0.00%	0.00%	0.00%	0.10%	0.02%	0.00%	0.00%	0.00%	0.00%	0.10%	0.02%	0.02%	0.02%	0.02%	0.02%	
120代		0.00%	0.00%	0.00%	0.00%	0.10%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	
.		62.00%	60.40%	66.40%	60.80%	55.50%	61.02%	65.20%	65.20%	69.30%	65.40%	65.40%	66.10%	63.56%	63.56%	63.56%		
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%	100.00%	100.00%	

データ理解 — 集計例

投稿者の傾向 (同行者別,目的-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		27.90%	16.40%	14.60%	45.40%	18.90%	24.64%	61.00%	65.90%	66.40%	58.30%	62.60%	62.84%	43.74%	
家族		57.30%	58.60%	61.20%	41.70%	64.60%	56.68%	25.10%	20.20%	19.40%	25.50%	25.30%	23.10%	39.89%	
恋人		6.60%	16.10%	15.30%	5.70%	9.20%	10.58%	6.50%	5.90%	7.80%	6.90%	4.70%	6.36%	8.47%	
友達		5.20%	7.50%	7.80%	4.40%	5.90%	6.16%	4.70%	3.80%	4.00%	6.90%	4.70%	4.82%	5.49%	
仕事仲間		1.90%	0.80%	0.40%	2.30%	0.50%	1.18%	1.60%	3.20%	1.10%	1.30%	2.20%	1.88%	1.53%	
その他		1.10%	0.60%	0.70%	0.50%	0.90%	0.76%	1.10%	1.00%	1.30%	1.10%	0.50%	1.00%	0.88%	
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%	

数値評価の構成 (評価値別, 目的-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		44.70%	53.80%	48.80%	53.30%	69.20%	53.96%	50.60%	45.30%	53.80%	52.40%	52.10%	50.84%	52.40%	
4		37.90%	32.90%	37.30%	35.20%	23.10%	33.28%	38.50%	39.90%	31.50%	36.60%	34.50%	36.20%	34.74%	
3		10.20%	7.70%	8.30%	7.30%	4.90%	7.68%	7.40%	9.80%	9.60%	7.50%	8.50%	8.56%	8.12%	
2		4.50%	3.90%	3.40%	2.40%	1.90%	3.22%	2.10%	3.30%	2.80%	2.20%	3.30%	2.74%	2.98%	
1		2.70%	1.70%	2.20%	1.80%	0.90%	1.86%	1.40%	1.70%	2.30%	1.30%	1.60%	1.66%	1.76%	
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

数値評価の平均 (エリア別)

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	4.29	4.18	4.07	4.34	4.29	4.34
01_登別	4.08	4.20	3.96	3.87	4.33	4.13	4.17
02_草津	4.29	4.27	4.13	4.04	4.38	4.18	4.33
03_箱根	4.26	4.16	4.18	4.05	4.28	4.25	4.27
04_道後	4.26	4.42	4.21	4.10	4.17	4.29	4.36
05_湯布院	4.58	4.39	4.40	4.30	4.52	4.60	4.58
■ B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.16	4.32
06_札幌	4.17	4.42	4.26	4.07	3.96	4.15	4.35
07_名古屋	4.07	4.29	4.17	3.99	3.91	4.03	4.24
08_東京	4.13	4.43	4.20	4.04	3.88	4.21	4.32
09_大阪	4.16	4.42	4.24	4.06	3.97	4.17	4.37
10_福岡	4.17	4.43	4.25	4.08	3.97	4.25	4.32

数値評価の平均 (レジャー, ビジネス別)

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	4.29	4.18	4.07	4.34	4.29	4.34
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.16	4.32

データ理解 — 集計例

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	4.56	4.42	4.40	4.38	4.59	4.51	4.58
男性	4.54	4.40	4.41	4.40	4.60	4.53	4.61
女性	4.58	4.44	4.39	4.36	4.58	4.48	4.56
■ B_ビジネス	4.31	4.46	4.28	4.20	4.02	4.31	4.43
男性	4.18	4.45	4.24	4.11	3.91	4.25	4.39
女性	4.48	4.47	4.34	4.31	4.17	4.40	4.49

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	4.33	4.19	4.08	4.34	4.31	4.40
男性	4.27	4.31	4.15	4.05	4.32	4.29	4.38
女性	4.39	4.35	4.26	4.13	4.36	4.33	4.43
■ B_ビジネス	4.15	4.39	4.23	4.06	3.95	4.10	4.31
男性	4.06	4.34	4.18	3.99	3.89	4.01	4.27
女性	4.36	4.52	4.37	4.24	4.10	4.31	4.41

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	4.24	4.21	4.14	4.01	4.37	4.26	4.32
男性	4.19	4.17	4.10	3.97	4.33	4.19	4.28
女性	4.38	4.35	4.25	4.14	4.47	4.47	4.42
■ B_ビジネス	4.07	4.48	4.19	4.03	3.99	4.04	4.32
男性	4.05	4.45	4.20	4.01	3.97	4.02	4.33
女性	4.13	4.57	4.14	4.06	4.05	4.11	4.30

結果の整理

	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"> ・ 	

議論&発表

- 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について発表してください
- 時間配分
 - グループごとで議論（10分）
 - グループごとに発表（3分 × 8グループ）

数値評価で違いを見るのは難しい

- ユーザーの8割が4~5の評価,
1~2をつけない→本音が見えない

数値評価の平均 (エリア別)

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	4.29	4.18	4.07	4.34	4.29	4.34
01_登別	4.08	4.20	3.96	3.87	4.33	4.13	4.17
02_草津	4.29	4.27	4.13	4.04	4.38	4.18	4.33
03_箱根	4.26	4.16	4.18	4.05	4.28	4.25	4.27
04_道後	4.26	4.42	4.21	4.05	4.28	4.25	4.36
05_湯布院	4.58	4.39	4.40	4.05	4.28	4.25	4.58
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.32	4.32
06_札幌	4.17	4.42	4.26	4.07	3.96	4.15	4.35
07_名古屋	4.07	4.29	4.17	3.99	3.91	4.03	4.24
08_東京	4.13	4.43	4.20	4.04	3.88	4.21	4.32
09_大阪	4.16	4.42	4.20	4.04	3.88	4.17	4.37
10_福岡	4.17	4.43	4.20	4.04	3.94	4.25	4.32

- 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

数値評価の平均 (レジャー, ビジネス別)

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	4.29	4.18	4.07	4.34	4.29	4.34
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.16	4.32

関連研究

- ・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.

数値評価の平均 (レジャー, ビジネス別)

行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニティ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.29	4.29	4.18	4.07	4.34	4.29	4.34
B_ビジネス	4.14	4.40	4.22	4.05	3.94	4.16	4.32

- ・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

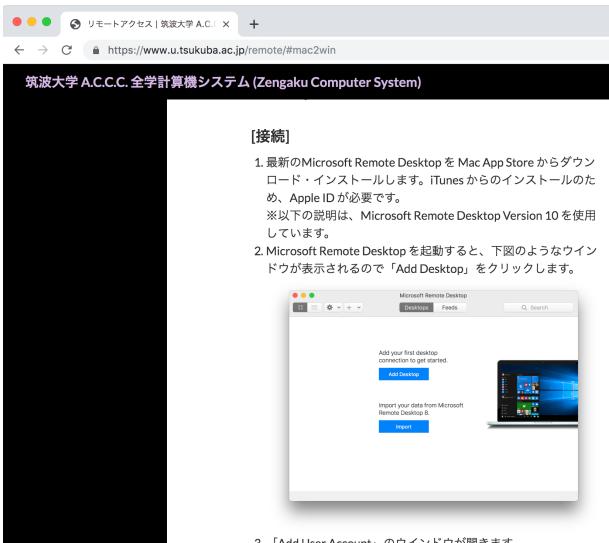
→ テキストと対応付ければ, 同じ点数でも差異があることを確認した

実習環境について

- 実習では,以下のツールを使用します
 - Part 2 では **Microsoft EXCEL** (用途: データの加工や修正)
 - Part 3 以降では **KHCoder** (用途: テキストマイニング) ※フリーソフト
- **KHCoder** は,全学計算機システムのリモートデスクトップでも動作します
 - 【Win】 <https://www.u.tsukuba.ac.jp/remote/#win2win>
 - 【Mac】 <https://www.u.tsukuba.ac.jp/remote/#mac2win>
- 個人のPCで **KHCoder** を使用しても構いません
 - ただし, Windows OS (11, 10, 8.1) を搭載した PC が必要です

全学計算機システムを利用する場合

- 事前に、下記のページの説明に従い全学計算機システム(Windows)へログインができるることを確認しておいてください
 - 【Win】 <https://www.u.tsukuba.ac.jp/remote/#win2win>
 - 【Mac】 <https://www.u.tsukuba.ac.jp/remote/#mac2win>



Mac の場合:

左記のページにある説明に従って、事前にツール [Microsoft Remote Desktop](#) のインストールが必要です

KH Coder インストール時の注意:

全学の Windows では C ドライブへのファイル保存は禁止されています。ダウンロードした KH Coder を解凍する場合は、保存先を「C ドライブ以外」に変更してください。例)「**Z:¥Desktop¥khcoder3**」

KH Coder のインストール (Part 3 以降で使用)

- ・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遷移先のページから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 保存先を「**Cドライブ以外**」(**Cドライブへの保存は禁止されています**)に変更します。例)
「Z:¥Desktop¥khcoder3」
- 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。

課題 — データをよく知る

- 以下の 2点を **PDF ファイルで提出** してください
 1. データ集計(ピボット分析)により作成した「集計表」(P.22~26)
 2. 「結果の整理」(P.27)
- 形式: PDF, 提出先: manaba, 期限: 次回～18:20)
- 個人 PC または全学リモートデスクトップに KHCoder のインストールを済ませておいてください ※ 提出物なし

環境準備 + Q&A

参考書

(KH Coder)

- [1] 樋口耕一. 社会調査のための計量テキスト分析—内容分析の継承と発展を目指して【第2版】 KH Coder オフィシャルブック. ナカニシヤ出版, 2020.
- [2] 樋口耕一. テキスト型データの計量的分析 —2つのアプローチの峻別と統合一. 理論と方法, 数理社会学会, 2004, 19(1): 101-115.
- [3] 牛澤賢二. やってみよう テキストマイニング —自由回答アンケートの分析に挑戦!. 朝倉書店, 2019
- New** [4] 樋口耕一. 動かして学ぶ! はじめてのテキストマイニング: フリー・ソフトウェアを用いた自由記述の計量テキスト分析 KH Coder オフィシャルブック II.ナカニシヤ出版, 2022.

(Windows環境によるデータ収集方法の参考に)

- [5] テキストマイニングソフトを利用した新未来洞察手法の研究. 第10分科会, (財)市場創造研究会. http://www.shijo-sozo.org/news/第10分科会_1.pdf

参考書

(Rを使った参考書)

- [6] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [7] 石田基広. "RMeCabによるテキスト解析. Rによるテキストマイニング入門." 森北出版, 2008, 51-82.

(他のツールを使った参考書)

- [8] 那須川哲哉. "テキストマイニングを使う技術/作る技術: 基礎技術と適用事例から導く本質と活用法." 東京電機大学出版局, 2006.
- [9] 上田隆穂, 黒岩祥太, 戸谷圭子. "テキストマイニングによるマーケティング調査." 講談社, 2005.

(統計解析を中心とした参考書)

- [10] 前田忠彦; 山崎誠. 言語研究のための統計入門. くろしお出版株式会社, 東京, 2013.