

# テキストマイニングの実践

## — 1日目 —

2021/6/25

人文社会ビジネス科学学術院  
ビジネス科学研究群

# 講義資料やデータの公開サイト

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

# スケジュール

- 1日目: 6/25(金)
    - 説明 — テキストマイニングの手順
    - 説明 — データをよく知る (Excel)
  - 2日目: 7/2(金)
    - 説明 — テキストマイニングツールの使い方 (KHCoder)
  - 3日目: 7/9(金)
    - 説明 — データ分析の実践 (KHCoder)
    - 実習 — データ分析の実践 (KHCoder) ←—————
  - 4日目: 7/16(金)
    - Text Mining Studio 利用体験
  - 体育の日: 7/23(金)
  - 5日目: 7/30(金)
    - 発表 — データ分析の実践 (KHCoder)
- ↑  
※ 注意 ※  
3日目と5日目 は, Zoom のブレイクアウトルーム機能を使ったグループワークになります。

# 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も多い
  - ・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](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



OMO 旭川  
旅のテンションを上げる3タイプの客室

# ファミリー # 女子旅



OMO  
旅のテンションを上げる3つの新サービスを開始

# ローカルの魅力

星野リゾート  
ホームページより  
(URL はスライド下)

# クチコミサイトの例 楽天トラベル

- ・ホテルのクチコミ数: 1,165万件 ※年間約60~70万



## 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 (今回)  
※ 2021/5/27現在

**鴨川シーワールドホテル クチコミ**

[travel.rakuten.co.jp/HOTEL/2910/review.html](http://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TOP > 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 Twitter Facebook ピン

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]

鴨川シーワールドホテル  
★★★★★ 4  
RENDEZ\_VOUSさんの鴨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。  
お部屋は古くて他の部屋の音(子供が走り回る足音など)が気になりました。  
畳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なのうれしいのですが、波の音が大きくて聞えて、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。土曜日にもう一度みてよかったです。  
シャチ以外のショーは平日でもしっかりしていました。

レポート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? 不適切なレビューを報告する はい いいえ

旅行の目的 … レジャー 同伴者 … 家族 宿泊年月 … 2015年06月

ご利用の宿泊プラン いい値! パリュープラン  
★鴨川シーワールドホテル (10畳バス・トイレ付) タイプ

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
- ★鴨シニュース★

**鴨川シーワールドホテル クチコミ**

[travel.rakuten.co.jp/HOTEL/2910/review.html](http://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

良いところ  
・部屋からの景色（朝日最高でした）  
・食事（品数多く、朝食とも良かったです）  
・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、それは思ひませんでした。  
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。

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。

【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リ付プラン  
【最安料金 (日安)】10,186円～ (消費税込11,000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 (日安)】10,278円～ (消費税込11,100円～)

【夜の水族館探検】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～)

【当月14:50からイルカと一緒にパリリ 2室限定】鴨川シーワールド体験付プラン  
【最安料金 (日安)】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大割引  
【5月・6月の月～木曜日限定プラン】  
【最安料金 (日安)】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★赤ちゃんが乗れるプラン  
【最安料金 (日安)】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 (日安)】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 (日安)】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 (日安)】11,389円～ (消費税込12,300円～)

【当日14:50～イルカ

旅行の目的 … レジャー 同伴者 … 家族 宿泊年月 … 2015年06月

ご利用の宿泊プラン 【洋室 禁煙・特別室】お部屋からシャチャやイルカも見える 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 07:33:49

鴨川シーワールドホテル 2015年06月11日 19:25:48

夫、2歳半と5ヶ月の子どもの4人で宿泊しました。  
【立地】当たり前ですが鴨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。

詳細にご感想頂きまして、ありがとうございます。  
今後の参考にして頂きます。  
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て、とても嬉しく思います。  
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。

【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す。

【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、泡ソーパ、支えのあるイス)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。しかしお風呂がわざとしきないので、温泉を楽しむという雰囲気ではなく、錢湯のお湯が温泉という感じです。  
また、23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。

【サービス】受付スタッフの方々とともに親切、丁寧です。チェックアウト後に子どもの薬を冷蔵庫にいれておいて欲しいとダメ元で頼むと快く入

## 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

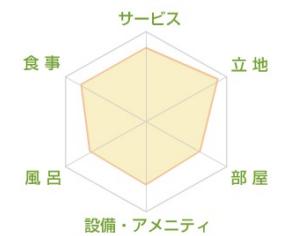
アンケート件数：886件

## 評価内訳

- 5点 ■■■■■ 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



## 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

## A. 楽天トラベル のクチコミデータ

- ・収集期間は **2018-2019** および **2020-2021(~5/12)** の **2セット**
- ・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

## B. ハッシュタグ「#新型コロナ」で投稿されたツイート

- ・収集期間は **2020-04-24 ~ 2021-05-05** の間で **1セット**
- ・Twitter APIで収集した **1%** サンプリングデータから **RTや重複を除去**
- ・データ件数は **32万件** からランダムサンプリングした **1万件**

# A. クチコミデータ

- ・楽天トラベルから収集した「お客様の声」のデータ
  - ・宿泊日が **2018-2019年** および **2020-2021年** (~5/12), 下記10エリア

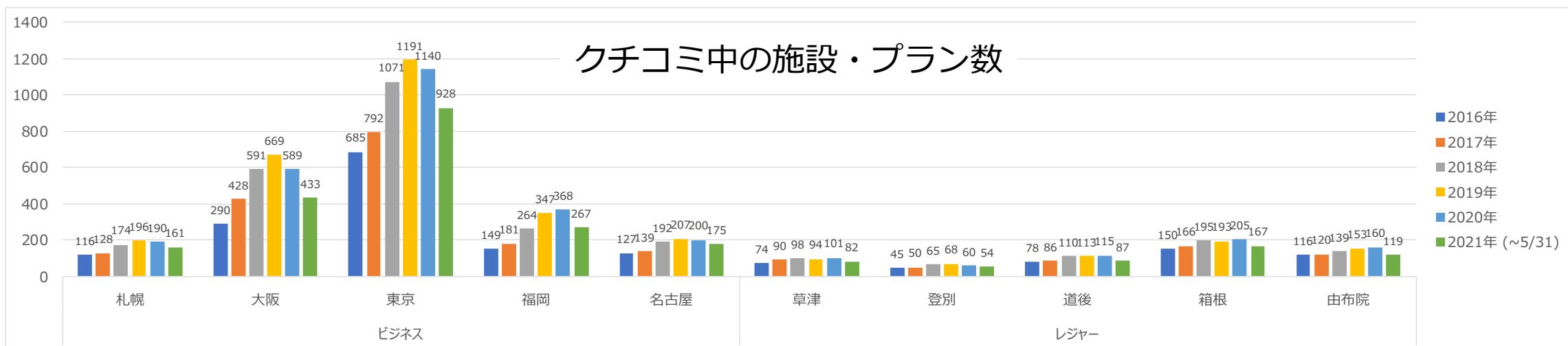
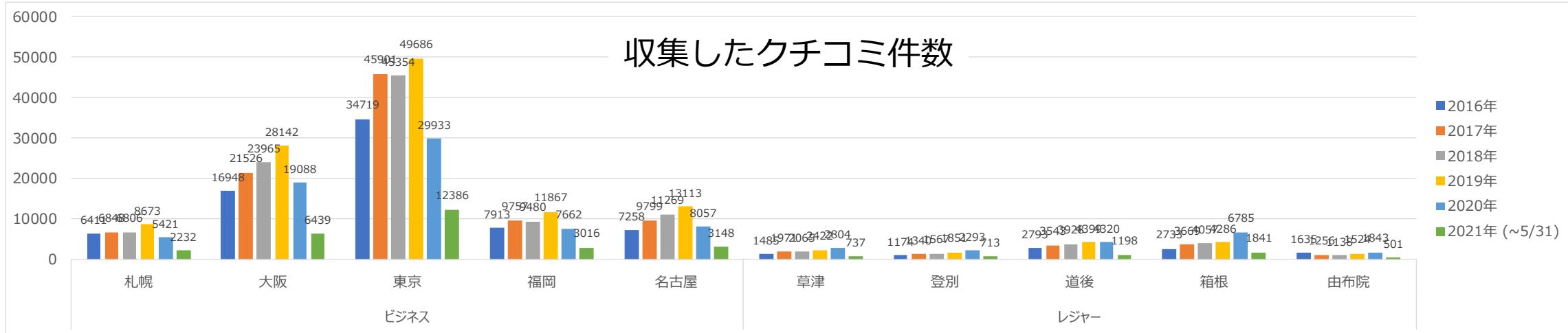
レジャー	5エリア	登別, 草津, 箱根, 道後, 湯布院	<b>1,000件</b> × 10エリア = 計10,000件
ビジネス	5エリア	札幌, 名古屋, 東京, 大阪, 福岡	

- ・データ項目

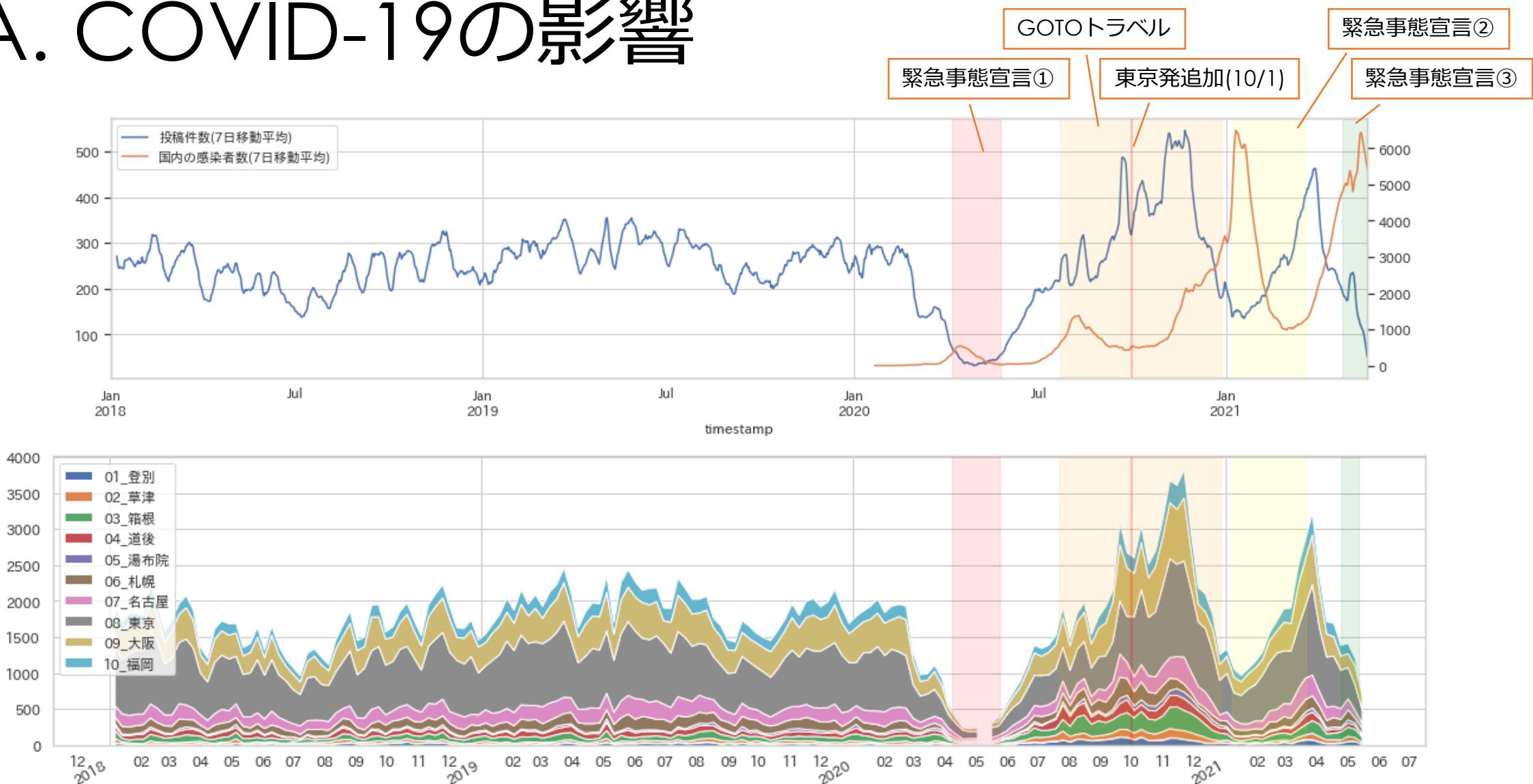
施設情報	4項目	カテゴリ, エリア, 施設番号, 施設名
口コミ	1項目	コメント
ユーザー評価	7項目	総合, サービス, 立地, 部屋, 設備・アメニティ, 風呂, 食事
その他の分類	2項目	旅行の目的, 同伴者
宿泊日	1項目	宿泊年月
ユーザー情報	3項目	ユーザー, 年代, 性別

# A. 全収集データの推移

※実習では2018~2021年のエリアごとに  
サンプリングした1,000件を使用



# A. COVID-19の影響

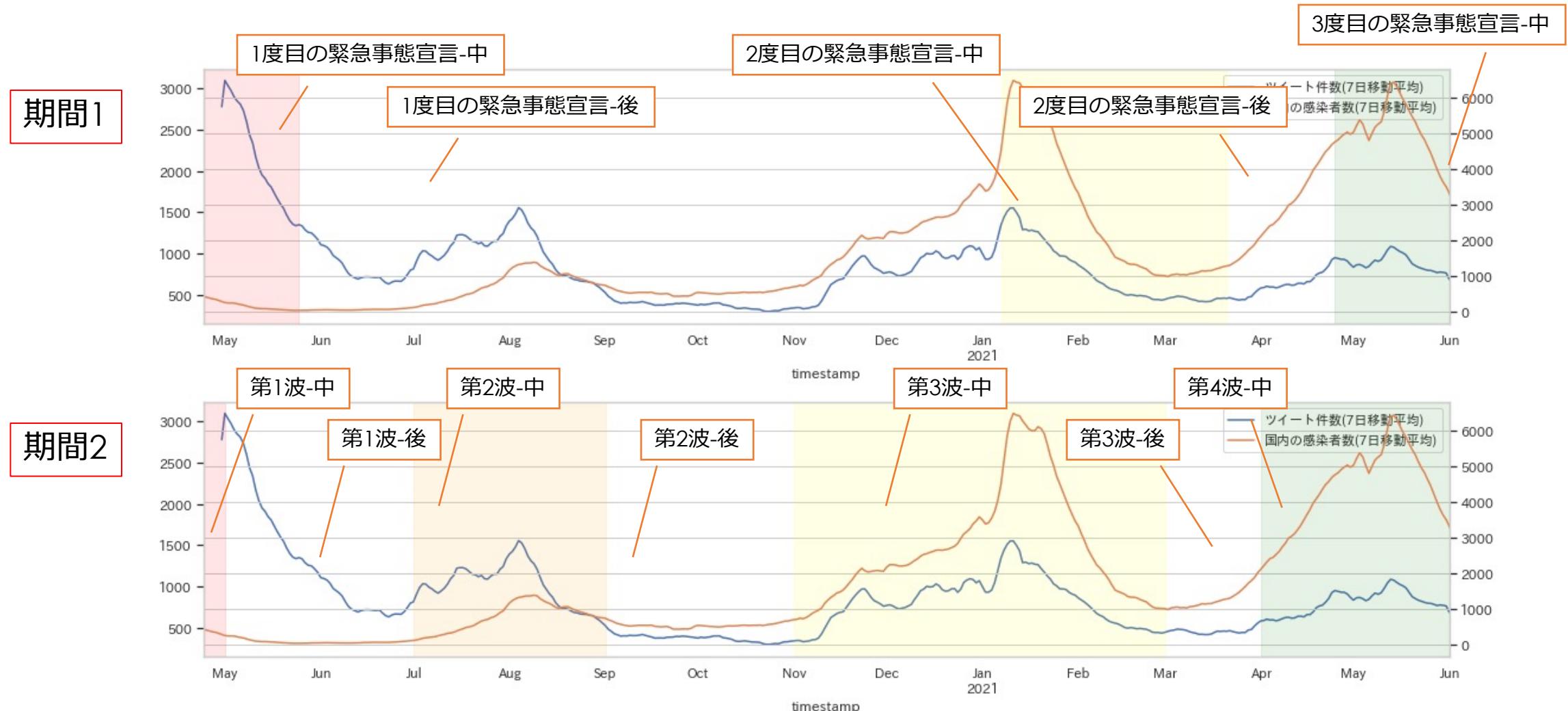


## B. ツイート

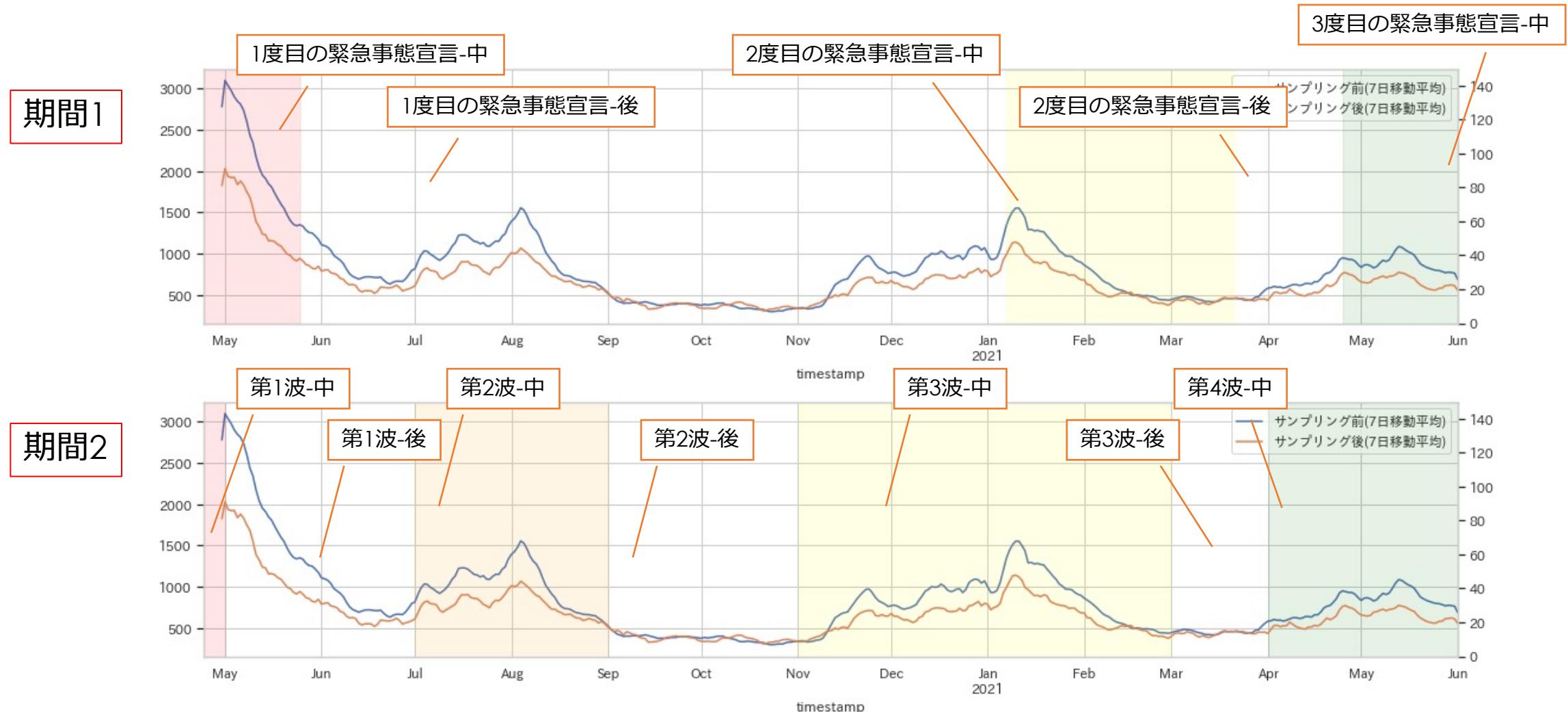
- ・ハッシュタグ「#新型コロナ」で投稿されたツイート
  - ・収集期間は **2020-04-24 ~ 2021-05-05** の 32万→**1万件をサンプリング**
  - ・データ項目

投稿日時	ツイート情報 5項目	ユーザ情報 4項目
ツイートの内容		
お気に入り数		
リツート数		
言語		
期間1	追加の属性情報 2項目	{1,2,3,4}度目の緊急事態宣言-{中,後}
期間2		第{1,2,3}波-{中,後}

## B. 追加した属性情報



# B. サンプリング前後の比較



# データ一覧

データファイル名	件数	データセット	備考
rakuten-1000-2020-2021.xlsx	10,000	<ul style="list-style-type: none"><li>レジャー+ビジネスの 10エリア</li><li>エリアごと 1,000件 (ランダムサンプリング)</li><li>期間: 2020/1/1~2021/5/12</li></ul>	<ul style="list-style-type: none"><li>本講義の全体を通して利用する</li></ul>
rakuten-1000-2018-2019.xlsx	10,000	<ul style="list-style-type: none"><li>レジャー+ビジネスの 10エリア</li><li>エリアごと 1,000件 (ランダムサンプリング)</li><li>期間: 2018/1/1~2019/12/31</li></ul>	<ul style="list-style-type: none"><li>実習用 (3~4日目)</li></ul>
covid19-10000.xlsx	10,000	<ul style="list-style-type: none"><li>ハッシュタグ「#新型コロナ」がついたツイート</li><li>Search API (1%サンプリング) で取得した 32万 →10,000件 (ランダムサンプリング)</li><li>期間: 2020/4/24~2021/5/31</li></ul>	<ul style="list-style-type: none"><li>実習用 (3~4日目)</li></ul>

# データの取得方法

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

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 hr

..

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		

本講義で主として使用 5/31

# ダウンロード方法



# データをよく知る

- ・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

# データをよく知る – 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 / コメント	列ラベル	男性	女性	na	総計
10代			0.00%	0.02%	0.00%	0.02%
20代			1.12%	1.17%	0.00%	2.29%
30代			2.15%	2.44%	0.00%	4.59%
40代			6.06%	3.82%	0.00%	9.88%
50代			9.12%	3.69%	0.00%	12.82%
60代			4.21%	1.12%	0.00%	5.33%
70代			0.81%	0.17%	0.00%	0.98%
80代			0.04%	0.02%	0.00%	0.06%
na			0.00%	0.00%	64.03%	64.03%
総計			23.52%	12.45%	64.03%	100.00%

投稿者の傾向 (エリア別)

行ラベル	A_レジャー	B_ビジネス	総計
男性	20.42%	26.62%	23.52%
女性	15.92%	9.04%	12.48%
na	63.66%	64.34%	64.00%
総計	100.00%	100.00%	100.00%

- 男性の投稿者が多い (女性の倍程度) → 男性の観点によるコメントが多い

- 
- 無回答(na)の中の分布が、表明した層と異なる(ある年代や性別に偏っている)可能性もある

# データをよく知る – 集計例

- 男女差は、レジャーに比べビジネスが大きい
- 男女差がレジャーで大きいのは道後

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
行ラベル		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	B_ビジネス 集計	
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%

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
行ラベル		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	B_ビジネス 集計	
10代		0.10%	0.00%	0.00%	0.00%	0.10%	0.04%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%
20代		1.90%	3.10%	2.91%	1.60%	3.40%	2.58%	1.30%	1.70%	1.30%	3.00%	2.70%	2.00%	2.29%
30代		5.90%	4.70%	5.21%	4.90%	5.80%	5.30%	4.61%	3.80%	3.70%	4.30%	3.00%	3.88%	4.59%
40代		11.50%	8.90%	7.52%	10.60%	8.00%	9.30%	9.62%	13.20%	9.40%	9.20%	10.90%	10.46%	9.88%
50代		10.00%	12.10%	10.02%	12.10%	13.50%	11.54%	15.73%	13.80%	13.60%	13.90%	13.40%	14.09%	12.82%
60代		5.40%	6.00%	5.51%	7.00%	6.40%	6.06%	5.21%	3.50%	4.90%	4.70%	4.70%	4.60%	5.33%
70代		1.10%	2.60%	1.40%	0.90%	1.10%	1.42%	0.90%	0.50%	0.40%	0.20%	0.70%	0.54%	0.98%
80代		0.10%	0.00%	0.10%	0.10%	0.00%	0.06%	0.00%	0.10%	0.20%	0.00%	0.00%	0.06%	0.06%
na		64.00%	62.60%	67.33%	62.80%	61.70%	63.69%	62.63%	63.40%	66.50%	64.70%	64.60%	64.37%	64.03%
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%

• 年代別では、目的によらず40~50代が多い

• あくまでも投稿者の傾向であって、旅行者の実態と一致するは限らない

# データをよく知る－集計例

## 投稿者の傾向 (同行者別) レジヤー

- レジヤーの中で一人が多いのは道後→道後は もはや仕事で行く場所 (性別でも男性が多い)

- レジヤーは家族が多く、ビジネスは一人が多い→出張は複数より単独が多い

個数 / コメント	列ラベル	A_レジヤー 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		21.80%	17.40%	14.40%	44.60%	14.10%	22.45%	64.40%	70.80%	67.50%	60.20%	62.40%	65.06%	43.76%		
家族		63.00%	60.90%	62.00%	42.00%	63.80%	58.34%	23.80%	17.80%	17.90%	23.40%	24.80%	21.54%	39.94%		
恋人		8.40%	13.10%	14.30%	6.00%	13.40%	11.04%	5.30%	6.30%	7.70%	7.20%	6.00%	6.50%	8.77%		
友達		5.00%	7.00%	7.50%	5.00%	7.20%	6.34%	3.80%	2.40%	4.80%	6.10%	4.30%	4.28%	5.31%		
仕事仲間		1.20%	0.80%	0.50%	2.00%	0.50%	1.00%	1.90%	2.20%	1.70%	2.20%	1.90%	1.98%	1.49%		
その他		0.60%	0.80%	1.30%	0.40%	1.00%	0.82%	0.80%	0.50%	0.40%	0.90%	0.60%	0.64%	0.73%		
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%	

## 数値評価の構成 (評価)

- 数値評価は、目的によらず高め→好評価しか投稿しない偏りがあるの可能性にも注意

- 高評価は、レジヤーがビジネスよりやや多い

個数 / コメント	列ラベル	A_レジヤー 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		47.20%	49.90%	49.10%	48.80%	68.70%	52.74%	51.20%	44.10%	45.70%	53.20%	48.30%	48.50%	50.62%		
4		34.20%	35.20%	31.20%	38.00%	21.70%	32.06%	35.90%	42.90%	39.10%	34.30%	36.80%	37.80%	34.93%		
3		10.70%	8.50%	10.30%	9.20%	4.50%	8.64%	8.80%	8.60%	9.40%	8.80%	8.90%	8.90%	8.77%		
2		4.90%	4.10%	6.00%	2.50%	3.20%	4.14%	2.40%	2.70%	3.20%	2.70%	3.80%	2.96%	3.55%		
1		3.00%	2.30%	3.40%	1.50%	1.90%	2.42%	1.70%	1.70%	2.60%	1.00%	2.20%	1.84%	2.13%		
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%	

- レジヤーの高評価は、湯布院が多く、登別がない

- ビジネスの高評価は、大阪と札幌が多く、東京都と名古屋がやや少ないが、僅差

# データをよく知る－集計例

数値評価の平均 (エリア別)

- レジャーは、風呂や食事が設備や部屋に比べて高評価

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	4.28	4.12	4.05	4.32	4.24	4.29
01_登別	4.13	4.24	3.97	3.96	4.38	4.10	4.18
02_草津	4.21	4.32	4.02	3.97	4.34	4.13	4.26
03_箱根	4.17	4.15	4.11	3.97	4.18	4.22	4.17
04_道後	4.18	4.38	4.14	4.02	4.14	4.20	4.30
05_湯布院	4.52	4.32	4.39	4.30	4.55	4.54	4.52
B_ビジネス	4.09	4.38	4.20	4.03	3.91	4.00	4.08
06_札幌	4.16	4.38	4.22	4.10	3.95	4.00	4.08
07_名古屋	4.10	4.31	4.16	3.98	3.89	4.00	4.08
08_東京	4.01	4.37	4.11	3.98	3.87	4.02	4.22
09_大阪	4.12	4.43	4.27	4.08	3.97	4.07	4.36
10_福岡	4.07	4.39	4.20	4.03	3.91	4.04	4.25

数値評価の平均 (レジャー, ビジネス別)

- レジャーもビジネスも立地が評価される
- ビジネスは、立地がその他に比べて高評価

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	4.28	4.12	4.05	4.32	4.24	4.29
B_ビジネス	4.09	4.38	4.20	4.03	3.91	4.04	4.28

# データをよく知る－まとめ

	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"> <li>約60%が年代や性別を表明していない</li> <li>年代別では、目的によらず40~60代が多い</li> <li>全体的に男性の投稿者が多い（女性の倍程度）</li> <li>レジャーに比べてビジネス方が男女差が大きい</li> <li>レジャーの中でも男女差が大きいのは道後</li> </ul>	<ul style="list-style-type: none"> <li>レビュー観点がある年代や性別に偏っている可能性</li> <li>無回答(na)中が、ある年代や性別に偏っている可能性</li> </ul>
目的別	<ul style="list-style-type: none"> <li>レジャーは家族が多い、ビジネスは一人が多い（出張は単独）</li> <li>レジャーの中でも、道後は男性の一人客が多い（道後はもはや仕事で行く場所）</li> </ul>	<ul style="list-style-type: none"> <li>レビューの観点が性別によって偏っている可能性</li> <li>レビューの観点がカテゴリと一致していない可能性（道後→仕事）</li> </ul>
数値評価 (総合)	<ul style="list-style-type: none"> <li>旅行目的によらず評価は高め</li> <li>レジャーがビジネスより評価が高め</li> <li>レジャーの中で高評価が多いのは湯布院、少ないのは登別</li> <li>ビジネスの中で高評価が多いのは大阪と札幌、少ないのは東京都と名古屋だが僅差</li> </ul>	<ul style="list-style-type: none"> <li>好評価しか投稿しない→コメントが好評価に偏っている可能性</li> <li>旅行目的によって投稿の動機が異なっている可能性</li> </ul>
数値評価 (項目ごと)	<ul style="list-style-type: none"> <li>レジャーの評価は、風呂や食事 &gt; 設備や部屋</li> <li>ビジネスの評価は、立地 &gt; その他</li> <li>レジャーの中で湯布院は軒並み高評価</li> <li>レジャーもビジネスも立地は高評価</li> </ul>	<ul style="list-style-type: none"> <li>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</li> </ul>
全体	<ul style="list-style-type: none"> <li>あくまでも楽天トラベルの特性であるので、旅行者の傾向として主張するためには別途裏付けが必要</li> </ul>	

# 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	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.24	4.28	4.12	4.05	4.32	4.24	4.29
B_ビジネス	4.09	4.38	4.20	4.03	3.91	4.04	4.28

- ・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

# 数値評価で違いを見るのは難

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	4.24	4.28	4.12	4.05	4.32	4.24	4.29
01_登別	4.13	4.24	3.97	3.96	4.38	4.10	4.18
02_草津	4.21	4.32	4.02	3.97	4.34	4.13	4.26
03_箱根	4.17	4.15	4.11	3.97	4.18	4.22	4.17
04_道後	4.18	4.38	4.14	4.03	4.30		
05_湯布院	4.52	4.32	4.39	4.03	4.52		
B_ビジネス	4.09	4.38	4.20	4.03	3.91		4.28
06_札幌	4.16	4.38	4.22	4.10	3.95	4.07	4.33
07_名古屋	4.10	4.31	4.16	3.98	3.89	3.96	4.25
08_東京	4.01	4.37	4.11	3.99	3.97	4.02	4.22
09_大阪	4.12	4.43	4.11	4.03	4.07	4.07	4.36
10_福岡	4.07	4.39	4.11	4.03	4.09	4.09	4.25

数値評価の平均 (レジャー, ビジネス別)

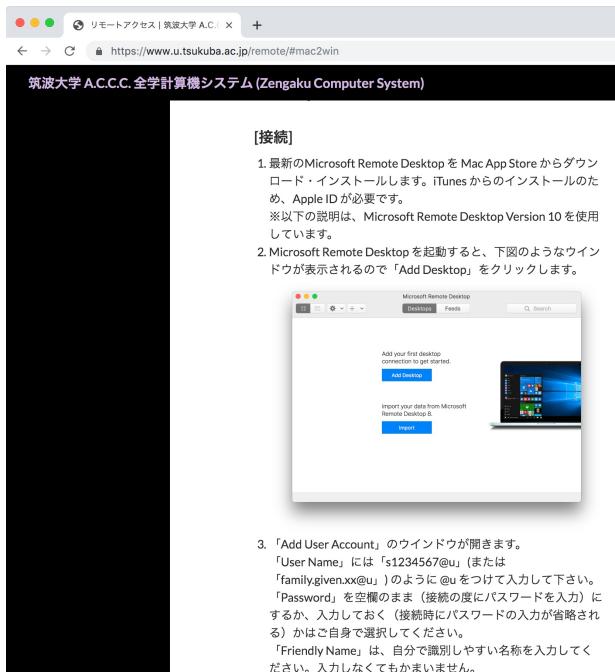
行ラベル	平均 / サービス	平均 / 立地	平均 / 部屋	平均 / 設備・アメニ	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.24	4.28	4.12	4.05	4.32	4.24	4.29
B_ビジネス	4.09	4.38	4.20	4.03	3.91	4.04	4.28

# 実習環境について

- ・実習では,全学計算機システムのリモートデスクトップを使用します
  - ・【Win】 <https://www.u.tsukuba.ac.jp/remote/#win2win>
  - ・【Mac】 <https://www.u.tsukuba.ac.jp/remote/#mac2win>
- ・個人のPCを使用しても構いません
  - ・ただし, Windows OS (10, 8.1, 7) を搭載した PC が必要です
  - ・以下のツールが使用できること確認してください
- ・実習では,以下のツールを使用します
  - ・次回: **Microsoft EXCEL** (用途: データの加工や修正)
  - ・次々回以降: **KHCoder** (用途: テキストマイニング) ※フリーソフト

# 全学計算機システムのリモートデスクトップ

- ・全学計算機システムのリモートデスクトップを使用します
  - ・【Win】 <https://www.u.tsukuba.ac.jp/remote/#win2win>
  - ・【Mac】 <https://www.u.tsukuba.ac.jp/remote/#mac2win>



上記のページにある説明に従って、全学計算機システム(**Windows**)へログインができますことを確認してください

## Mac の場合:

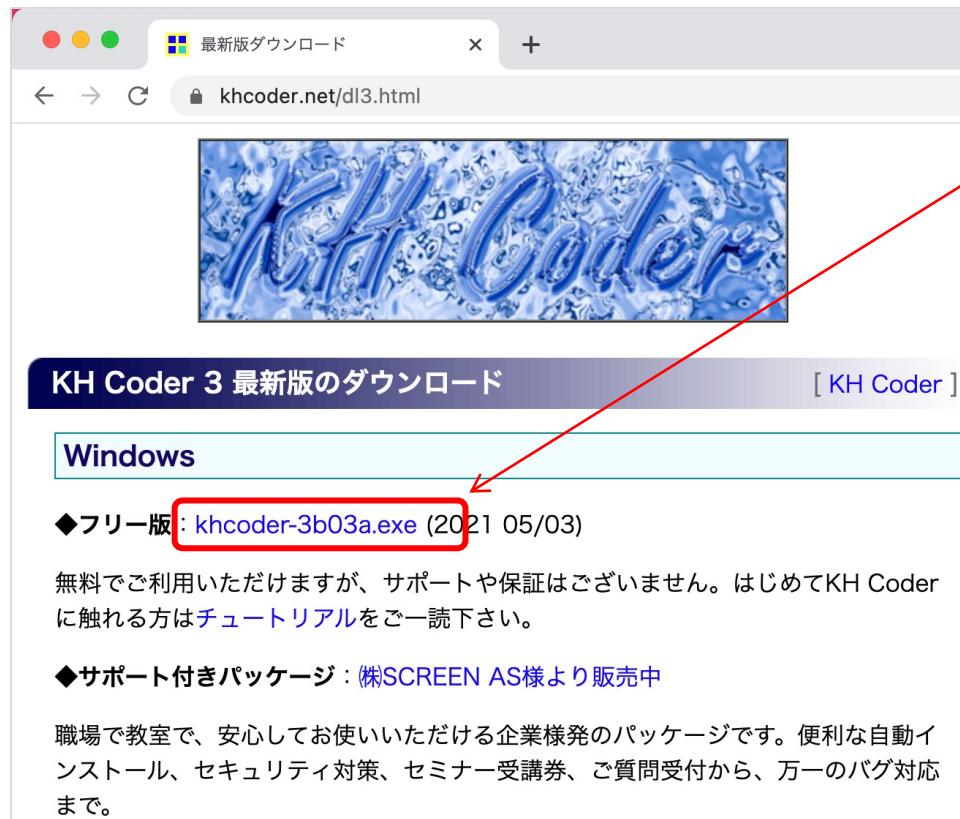
左記のページにある説明に従って、事前にツール **Microsoft Remote Desktop** のインストールが必要です

## KH Coder インストール時の注意:

全学の Windows の場合は、ログイン後の**デスクトップ上に「khcoder」というフォルダを作成**して、その中に解凍してください ※Cドライブへの保存は禁止されています

# KH Coder のインストール (次回の実習で使用)

- ・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

- EXCELを使って、P.24~27 のデータ集計(ピボット分析)を行い、作成したシートを PDFファイルで提出してください

サンプル:

[https://github.com/haradatm/lecture/blob/master/gssm-202107/03-samples/practice-1\\_sample.xlsx](https://github.com/haradatm/lecture/blob/master/gssm-202107/03-samples/practice-1_sample.xlsx)

- 形式: PDF, 提出先: manaba, 期限: 次回 7/2 21:00

# 環境準備 + Q&A

# 参考書

## (KH Coder)

- [1] 横口耕一. 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] 横口耕一. 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—. 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

## (Windows環境によるデータ収集方法の参考に)

- [4] 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](http://www.shijo-sozo.org/news/第10分科会_1.pdf)

# 参考書

## (Rを使った参考書)

- [5] 金明哲. "テキストデータの統計科学入門." 岩波書店, 2009.
- [6] 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)

- [7] 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.
- [8] 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)

- [9] 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.