テキストマイニングの実践
― 1日目 ―

2020/7/1

ビジネス科学研究科
経営システム科学専攻

講義資料やデータの公開サイト

https://github.com/haradatm/lecture/tree/master/**gssm-202007**

講義スライド

- https://github.com/haradatm/lecture/tree/master/**gssm-202007**

スケジュール

- 1日目: 7/1(水)
  - 説明 — データ分析の手順
  - 実習 — 環境構築, データをよく知る (Excel)
- 2日目: 7/10(金)
  - 説明 — テキストマイニング ツールの使い方 (KHCoder)
  - 説明 — データ分析の実践 (KHCoder)
- 3日目: 7/17(金)
  - 外部講師 — NTTデータ数理システム
- 体育の日: 7/24(金)
- 4日目: 7/31(金)
  - 説明 — データ分析の実践 (KHCoder)
  - 実習 — データ分析の実践 (KHCoder)
- 5日目: 8/7(金)
  - 発表 — データ分析の実践 (KHCoder)

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, レビューサイトの口コミ, Twitter 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 — コックローチ

- 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

星野リゾート
ホームページより
(URL はスライド下)

- 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

https://travel.watch.impress.co.jp/docs/news/1056715.html
https://www.hoshinoresorts.com/brand/omo/

口コミサイトの例

- ホテルの口コミ数: 1,098万件 ※年間約60～70万 →今回50万件

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 (今回)

※ 2020/5/30現在

鴨川シーワールドホテル クチ

HARADA Tomohiko

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

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　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　懸賞広場　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]

鴨川シーワールドホテル

癒しめぐりの宿　夢みさき

もっと見る

施設関連情報

- 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★

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

ホテル・旅行のクチコミTOP

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

投稿の種類：クチコミ（感想・情報）　クチコミ（苦情）

宿泊年月：指定なし（7

キーワード

同伴者：一人　家族　恋人　友達　仕事仲間

年代：指定なし

性別：指定なし　男性　女性

絞り込む　絞り込みを解除

クチコミを投稿する

クチコミを修正する

[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 鴨川シーワールドホテル のクチコミ (感想・情報)

RENDEZ_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。
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はうれしいのですが、
波の音が大きく聞こえて、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。土曜日にもう一度みてよかった。
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　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？　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！バリュープラン♪
ご利用のお部　【wa海側和室（10畳バス・トイレ付）タイプ】

宿泊プラン一覧

- 【1泊朝食付＆カモシー入園パス付】朝からカモシーへＬｅｔ's Ｇｏ！ ［最安料金（目安）］7,963円～（消費税込8,600円～）
- 【1日パスポート/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】カモシー入園なし　ビジネスプラン ［最安料金（目安）］9,352円～（消費税込10,100円～）
- 【カモシー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～

鴨川シーワールドホテル クチ

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

HARADA Tomohiko

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

外国語サイト

- Book Kamogawa Sea World Hotel
- Hotels in Sotobo(Kamogawa, Katsuura, Onjuku, Mobara)
- KAMOGAWA SEA WORLD HOTEL 預訂
- 外房(鴨川•勝浦•御宿•茂原) 酒店一覧

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（品数も多く、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、そう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です。

レビューを評価してください　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？　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

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りますので、
お客様からの声として、
スタッフと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。

ご利用の宿泊プラン　【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
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

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。

【部屋】至って普通です。(古いからか、隣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。清浄機などもTEL一本ですぐに届けて下さいました。

【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。

【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、泡ソープ、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。しかしお風呂がひとつしかないので、温泉を楽しむという雰囲気ではなく、銭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れ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　2015年06月11日 19:25:48

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。

詳細にご感想頂きまして、ありがとうございます。
今後の参考にさせて頂きます。
また、スタッフ対応に関しまして、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、
スタッフと共有させて頂きます。

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。
早急に対応をして、改善を行います。
貴重なご意見を、ありがとうございます。

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チリ付プラン　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チリ　2室限定】鴨川シーワールド体験付プラン ［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＆シーワールド入園パ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★赤ちゃんなっ得プラン ［最安料金（目安）］10,926円～（消費税込11,800円～）
- お子様にも大好評！オーシャンビュープラン ［最安料金（目安）］11,112円～（消費税込12,000円～）
- 【80cmのジャンボサイズ】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～）
- 房総2大テーマパーク満喫「マザー牧場チケット」付プラン ［最安料金（目安）］11,389円～（消費税込12,300円～）
- 【当日14：50～イルカ

テキストマイニング

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

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・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（品数も多く、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、そう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です。

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した。

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
お客様からの声として、
スタッフと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。

テキストデータ

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

実習で使用するデータ

- 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2019年,下記の10エリアが対象

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	

  - 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

参考 ― データ収集の推移

※演習では 2019年のエリアごとに
サンプリングした1,000件を使用

参考 ― サンプリング状況

※演習では 2019年のエリアごとに
サンプリングした1,000件を使用

NO.	カテゴリ	エリア	クチコミ件数					全施設・プラン数				
			2016年全件	2017年全件	2018年全件	2019年全件	カバー率	2016年全件	2017年全件	2018年全件	2019年全件	カバー率
1	レジャー	登別	1,174	1,340	1,567	1,851	54.02%	45	50	65	68	92.65%
2		草津	1,485	1,971	2,065	2,422	41.29%	74	90	98	94	93.62%
3		箱根	2,733	3,669	4,057	4,286	23.33%	150	166	195	193	81.35%
4		道後	2,793	3,543	3,928	4,399	22.73%	78	86	110	113	84.07%
5		由布院	1,635	1,256	1,138	1,524	65.62%	116	120	139	153	88.89%
6	ビジネス	札幌	6,411	6,848	6,806	8,673	11.53%	116	128	174	196	77.04%
7		名古屋	7,258	9,799	11,269	13,113	7.63%	127	139	192	207	78.26%
8		東京	34,719	45,901	45,354	49,686	2.01%	685	792	1,071	1,191	41.39%
9		大阪	16,948	21,526	23,965	28,142	3.55%	290	428	591	669	47.38%
10		福岡	7,913	9,757	9,480	11,867	8.43%	149	181	264	347	57.93%
全体			83,069	105,610	109,629	125,963	7.94%	1,830	2,180	2,899	3,231	57.66%

ダウンロードできるデータ

データファイル名	件数	データセット	備考
rakuten_2019.xlsx	10,000	• 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) • EXCEL 形式 (シート名「2019」)	• 本講義の全体を通して利用する
rakuten_2020.xlsx	8,518	• レジャー+ビジネスの 10エリア • エリアごと 1,000件 (登別,草津,由布院は1,000件以下のため全数, それ以外はランダムサンプリング) • EXCEL 形式 (シート名「2020年」)	• 演習用 (3~4日目)
corona_2020.xlsx	10,000	• 2020/4/27~5/30 のハッシュタグ「#新型コロナ」がついたツイート • Search API (1%サンプリング) で取得 • EXCEL 形式 (シート名「corona」)	• 演習用 (3~4日目)

データの取得方法

- https://github.com/haradatm/lecture/tree/master/gssm-202007

ダウンロード方法

- Download ボタンをクリックするとダウンロードを開始

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_2019.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)

個数 / コメン	列ラベル			
行ラベル	男性	女性	na	総計
10代	0.01%	0.03%	0.00%	0.04%
20代	0.75%	0.89%	0.00%	1.64%
30代	2.38%	2.39%	0.00%	4.77%
40代	6.78%	3.81%	0.00%	10.59%
50代	9.35%	3.24%	0.00%	12.60%
60代	4.83%	1.29%	0.00%	6.12%
70代	0.59%	0.24%	0.00%	0.83%
80代	0.07%	0.01%	0.00%	0.08%
na	0.00%	0.00%	63.33%	63.33%
総計	**24.77%**	**11.90%**	**63.33%**	**100.00%**

投稿者の傾向 (エリア別)

個数 / コメン	列ラベル		
行ラベル	A_レジャー	B_ビジネス	総計
男性	23.38%	26.20%	24.79%
女性	13.60%	10.22%	11.91%
na	63.02%	63.58%	63.30%
総計	**100.00%**	**100.00%**	**100.00%**

- 男性の投稿者が多い (女性の倍以上) →男性の観点によるコメントが多い

- 無回答(na)の層が,表明した層の分布と異なる(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					A_レジャー 集計	⊟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		06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
男性	24.50%	22.30%	17.70%	31.30%	21.10%	23.38%	29.20%	27.20%	23.90%	23.70%	27.00%	26.20%	24.79%
女性	13.90%	13.50%	15.60%	9.60%	15.40%	13.60%	8.70%	8.50%	12.40%	11.50%	10.00%	10.22%	11.91%
na	61.60%	64.20%	66.70%	59.10%	63.50%	63.02%	62.10%	64.30%	63.70%	64.80%	63.00%	63.58%	63.30%
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)

- 年代別では,目的によらず40～50代が多い

個数 / コメント	列ラベル												
	⊟A_レジャー					A_レジャー 集計	⊟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		06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
10代	0.00%	0.10%	0.00%	0.00%	0.00%	0.02%	0.10%	0.10%	0.00%	0.00%	0.10%	0.06%	0.04%
20代	0.70%	2.80%	2.90%	0.60%	3.00%	2.00%	1.20%	1.70%	1.20%	1.50%	0.80%	1.28%	1.64%
30代	5.10%	5.10%	7.10%	4.60%	6.10%	5.60%	3.90%	3.50%	3.70%	3.60%	5.00%	3.94%	4.77%
40代	12.50%	8.80%	7.20%	9.50%	9.30%	9.46%	11.40%	11.50%	10.70%	13.30%	11.70%	11.72%	10.59%
50代	12.60%	11.80%	8.80%	15.50%	10.40%	11.82%	14.10%	12.80%	14.30%	12.20%	13.40%	13.36%	12.59%
60代	6.90%	5.60%	5.60%	9.50%	6.60%	6.84%	6.00%	5.70%	5.90%	4.20%	5.20%	5.40%	6.12%
70代	0.60%	1.50%	1.60%	1.10%	1.00%	1.16%	1.00%	0.30%	0.30%	0.30%	0.60%	0.50%	0.83%
80代	0.00%	0.00%	0.10%	0.10%	0.00%	0.04%	0.10%	0.10%	0.20%	0.10%	0.10%	0.12%	0.08%
na	61.60%	64.20%	66.70%	59.10%	63.50%	63.02%	62.10%	64.30%	63.70%	64.80%	63.00%	63.58%	63.30%
110代	0.00%	0.10%	0.00%	0.00%	0.10%	0.04%	0.10%	0.00%	0.00%	0.00%	0.00%	0.02%	0.03%
90代	0.00%	0.00%	0.00%	0.00%	0.00%			0.00%	0.00%	0.00%	0.10%	0.02%	0.01%
総計	**100.00%**	**100.00%**	**100.00%**										

- あくまでも投稿者の傾向であって,一般旅行者の実態ではないことに注意

データをよく知る — 集計例

- レジャーの中で道後は一人が多い →道後はもはや仕事で行く場所 (性別でも男性が多い)

- レジャーは家族が多く，ビジネスは一人が多い →出張は複数より単独が多い

投稿者の傾向 (同行者別,目的

個数 / コメント	列ラベル												
	A_レジャー					A_レジャー 集計	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		06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
一人	26.00%	11.50%	12.80%	49.60%	12.40%	22.46%	69.00%	71.20%	73.50%	63.80%	65.60%	68.62%	45.54%
家族	61.10%	64.80%	63.50%	35.40%	65.30%	58.02%	22.30%	19.40%	17.60%	25.00%	22.10%	21.28%	39.65%
恋人	5.40%	13.90%	13.50%	3.30%	12.40%	9.70%	2.70%	2.80%	2.90%	3.10%	3.60%	3.02%	6.36%
友達	5.30%	8.20%	9.00%	6.90%	8.10%	7.50%	4.10%	3.80%	3.30%	5.20%	3.90%	4.06%	5.78%
仕事仲間	1.40%	1.20%	0.50%	3.70%	0.80%	1.52%	1.80%	2.30%	2.00%	2.40%	4.00%	2.50%	2.01%
その他	0.80%	0.40%	0.70%	1.10%	1.00%	0.80%	0.10%	0.50%	0.70%	0.50%	0.80%	0.52%	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%

- 数値評価は,目的によらず高め →好評価しか投稿しないバイアスの可能性にも注意

- 数値評価は,レジャーがビジネスより高い

数値評価の構成 (評価

個数 / コメント	列ラベル												
	A_レジャー					A_レジャー 集計	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		06_札幌	07_名古屋	08_東京	09_大阪	10_福岡		
5	37.80%	47.40%	47.10%	42.40%	69.70%	48.88%	36.00%	36.50%	35.50%	36.70%	32.70%	35.48%	42.18%
4	38.80%	36.60%	35.50%	39.60%	22.20%	34.54%	44.40%	46.20%	44.80%	46.90%	44.80%	45.42%	39.98%
3	13.50%	9.50%	9.60%	11.40%	5.10%	9.82%	14.50%	11.70%	11.70%	10.80%	14.60%	12.66%	11.24%
2	6.90%	3.30%	5.00%	3.60%	1.70%	4.10%	3.60%	3.80%	4.40%	3.70%	5.40%	4.18%	4.14%
1	3.00%	3.20%	2.80%	3.00%	1.30%	2.66%	1.50%	1.80%	3.60%	1.90%	2.50%	2.26%	2.46%
総計	100.00%	100.00	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	00.00%	100.00%	100.00%	100.00%

- レジャーの数値評価は,湯布院が高く,登別が低い

- ビジネスの数値評価は,大阪と名古屋が高く,東京都と福岡がやや低いが,僅差

データをよく知る — 集計例

数値評価の平均 (エリア別)

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	**4.16**	**4.20**	**4.04**	**3.98**	**4.22**	**4.22**	**4.23**
01_登別	3.88	4.13	3.83	3.78	4.16	3.93	4.02
02_草津	4.12	4.29	3.95	3.87	4.25	4.15	4.22
03_箱根	4.18	4.07	4.04	3.97	4.19	4.27	4.19
04_道後	4.01	4.21	3.97	3.90	3.96	4.14	4.15
05_湯布院	4.60	4.29	4.43	4.35	4.53	4.60	4.57
B_ビジネス	**3.91**	**4.24**	**3.98**	**3.85**	**3.69**	[illegible]	[illegible]
06_札幌	3.95	4.20	4.01	3.81	3.66	[illegible]	[illegible]
07_名古屋	3.96	4.15	3.99	3.89	3.74	[illegible]	[illegible]
08_東京	3.85	4.33	3.91	3.81	3.67	3.91	4.04
09_大阪	3.92	4.32	4.04	3.90	3.74	4.02	4.13
10_福岡	3.85	4.21	[illegible]	[illegible]	[illegible]	[illegible]	4.00

- レジャーは,風呂や食事が,設備や部屋に比べて高評価
- 湯布院は,レジャーの中で,軒並み高評価が多い
- 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

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	4.16	4.20	4.04	3.98	4.22	4.22	4.23
B_ビジネス	3.91	4.24	3.98	3.85	3.69	3.94	4.08

練習 — データをよく知る

- 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)してください

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?

まとめ方の一例

	データの特徴	注意すべきバイアス等
年代別・性別	・ 約60%が年代や性別を表明していない ・ 年代別では,目的によらず40～60代が多い ・ 全体的に男性の投稿者が多い (女性の倍以上) ・ レジャーに比べてビジネス方が男女差が大きい ・ レジャーの中でも男女差が大きいのは道後	・レビュー観点がある年代や性別に偏っている可能性 ・無回答(na)層が,ある年代や性別に偏っている可能性
目的別	・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) ・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)	・レビューの観点が性別によって偏っている可能性 ・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)
数値評価 (総合)	・旅行目的によらず評価は高め ・レジャーがビジネスより評価が高め ・レジャーの中で評価が高いのは湯布院,低いのは登別 ・ビジネスの中で評価が高いのは札幌と大阪, 低いのは東京都と福岡だが僅差	・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 ・旅行目的によって投稿の動機が異なっている可能性
数値評価 (項目ごと)	・レジャーの評価は, 風呂や食事 > 設備や部屋 ・ビジネスの評価は, 立地 > その他 ・レジャーの中で湯布院は軒並み高評価 ・レジャーもビジネスも立地は高評価	・旅行目的によって評価の観点が異なっている可能性

全体	・あくまでも楽天トラベルの特性であるので,旅行者一般として主張するには別途裏付けが必要

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	4.16	4.20	4.04	3.98	4.22	4.22	4.23
B_ビジネス	3.91	4.24	3.98	3.85	3.69	3.94	4.08

- 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

演習環境について

- 演習では,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 (7, 8.1, 10) 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

- 演習では,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)へログインができる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

KH Coder のインストール (次々回の演習で使用)

- 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 指示に従いインストール

自己解凍ファイルです。このファイルを実行（ダブルクリック）し、開いたWindowの「Unzip」ボタンをクリックすると、（特に変更しなければ）「C:¥khcoder3」という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。

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—. ナカニシヤ出版, 2014.

[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—. 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CGM収集の参考に)

[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

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.