

# テキストマイニングの実習

## －1日目－

2016/7/6

ビジネス科学研究科  
経営システム科学専攻

# スケジュール

- 7/6

- 説明 データ分析の手順
- 演習 データの理解 (Excel)

- 7/13

- 説明 ツール (KHCoder)
- 練習 ツール (KHCoder)

- 7/20

- 演習 データの分析 (KHCoder)

# テキストマイニング

- テキストから有益な情報を抽出する技術の総称  
「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
  - 昔から
    - 営業日報
    - 自由記述のアンケート
    - コールセンタの応対ログ
  - 近年 → CGM (コンシューマー・ジェネレイテッド・メディア) なども
    - レビューサイトの口コミ
    - ブログやマイクロブログ (Twitter, Facebook)

# CGMが注目される理由 (1/2)

- スマフォや SNS の普及 →  
SNS の口コミやレビューサイトの重要性が増加
- 商品購入時に参考とする情報・広告:
  - 購入サイト・レビューサイトの口コミが 47.9%
  - SNSでの口コミ: 17.2% (スマートフォン保有者 n=535)

出所: 総務省「ICTの進化がもたらす社会へのインパクトに関する調査研究」(平成26年)

- 影響を受けやすいSNSがある: 全体の 20.8%
  - 特に,若年層を中心にSNSの口コミが浸透
  - 10代女性は49.4%, 10代男性は33.7% で高い傾向 (各 n=83)

出所: 日本通信販売協会「ネット通販に関する消費者実態調査2013」

# CGMが注目される理由 (2/2)

- 日々,ネット上に製品やサービスに関する消費者の膨大な評価情報が蓄積

消費者: 有用な情報取得・共有ツール

企業: 消費者の評判に関する情報源



- 自社のビジネスに役立ちそうな情報の収集し,活用したい

# 口コミサイトの例



- ホテルの口コミ数: 900万件 ※年間約60万件増加
  - 経年変化: 780万件(一昨年)→836万件(昨年)→900万件

The screenshot shows the Rakuten Travel website's review section. At the top, it displays "お客様の声" (Customer Reviews) with a count of "ホテルのクチコミ数No.1 9,000,641件". Below this, there are search filters for "国内宿泊" (Domestic Accommodation) and "海外宿泊" (Overseas Accommodation). A green banner at the bottom left says "新着! 最新のクチコミ" (Newest reviews) and lists recent reviews from June 2017, such as "ホテルユニゾ京都四条烏丸のクチコミ (186件)" with a 4.25 rating. On the right side, there is a sidebar with a QR code for mobile投稿 (posting) and a message encouraging users to share their experiences.

投稿日	ホテル名	評価
2017-06-02 23:58:16	ホテルユニゾ京都四条烏丸のクチコミ (186件)	★★★★★ 4.25
2017-06-02 23:56:41	東横イン湘南平塚駅北口1のクチコミ (499件)	★★★★★ 3.66
2017-06-02 23:55:35	ホテル京阪京都グランデのクチコミ (2813件)	★★★★★ 4.11
2017-06-02 23:55:23	輪島温泉 八汐のクチコミ (383件)	★★★★★ 4.31
2017-06-02 23:52:22	伊東温泉 伊東ホテル聚楽(じゅらく)のクチコミ (1284件)	★★★★★ 4.18
2017-06-02 23:51:44	温泉旅館 水月のクチコミ (19件)	★★★★★ 4.64
2017-06-02 23:50:25	湯けむりとにかくごり湯の宿 霧島国際ホテルのクチコミ (819件)	★★★★★ 3.99

# 口コミサイトの例

**R 横川シーウールドホテル クラ**

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

HARADA Tomohiko

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 スーパーDEAL 30%以上ポイントバック!

宿泊料金 旅行料金

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 観覧券内

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 > 横川シーウールドホテル クチコミ・感想・情報

鴨川シーウールドホテル

★★★★★ 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

問い合わせメール チュアモモ 3

日程からプランを探す

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

横川シーウ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  
4点  
3点  
2点  
1点

項目別評価  
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

外観  
立地  
部屋  
設備・アメニティ  
風呂  
食事

旅館の種類 クチコミ（感想・情報） クチコミ（苦情） 一人 家族・恋人 友達 仕事仲間

宿泊年月 指定なし(7) 年代 性別 指定なし 男性 女性 紹介込みを解除

キーワード

クチコミを投稿する

宿泊プラン一覧

【1泊朝食付カモシモー入園パック】朝からカモシモーでlet's Go! 【最安料金（目安）】7,963円～ (消費税込8,600円～)

【1泊バスポート朝食付+翌日のランチ券付】運んでいたもたっぷり楽しめ！ 【最安料金（目安）】8,159円～ (消費税込8,900円～)

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

【カモシモーオリジナルパンチング】6月のお誕生日限定★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～)

【シャツのイラスト入りオリジナルフィットスタイル付】7月のお誕生日限定★雨だってへっちゃら♪ 【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～

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？ 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？

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 ご利用のお部屋 [wa海側室 (10畳バス・トイレ付) タイプ]

**R 横川シーウールドホテル クラ**

[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 横川シーウールドホテル 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、そう思いました。

気にかかる事は色々ありましたが、フロントのお姉さんが一生懸命で、その笑顔に惚れました。

プロンプ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。

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】お部屋からシャツやイルカも見える シーウールドと海一宿泊プラン ご利用のお部屋 [wa5シーウールド] 見える特別室禁煙【洋室】

総合 ★★★★★ 4 投稿者さんの 横川シーウールドホテル 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】当たり前ですが横川シーウールドにて近くもあり、ゆっくりお風呂内を見学できました。

詳細にご感想頂きました、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、スタッフ対応にもしっかりとお褒めのお言葉を頂戴しまして、とても嬉しいです。

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。

【部屋】至って普通です。（古いからか、牀の声は少し聞こえます。）トイレ蛇足などはしっかりさわっていました。清淨便などTEL一本で届けて下さいました。

【食事】夜寝共にパンギング。イエスですが子ども用イース、エコノ、ベビーベッドも用意して下さいました。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き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(ペビーベッド、コーン、バス、おもちゃ、浴ソーブ、支えのあるバス)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に、温泉を楽しむという雰囲気ではなく、戻るの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もて親切、丁寧です。チェックアウト後に子どもの姿を冷蔵庫にいておいて欲しいとダメ元で探すと早く入

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ハーフトリッププラン 設定期定  
【最安料金（目安）】10,278円～ (消費税込11,100円～)  
【当日13:40～エコニアクリームコムユニケーション】トライアム 1日3組 設定期定  
【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ヶ月～10月の火・水曜日 設定期定  
【最安料金（目安）】10,278円～ (消費税込11,100円～)  
【当日4:50からイルカと一緒にパブリックラン】朝からカモシモーで大満足♪5月・6月の月～木曜日 設定期定  
【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&シーウールド入園パスポート付で大満足♪5月・6月の月～木曜日 設定期定  
【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】朝から泊りお泊りさんも嬉しい★赤ちゃん用グッズプラン  
【最安料金（目安）】10,926円～ (消費税込11,800円～)  
お子様にも大好評！オーシャンシーウーブラン  
【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】海の王者キャッチねぐらプラン  
【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「マザーフームチケット」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

# 口コミサイトの例



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  
4点  
3点  
2点  
1点

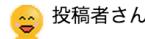
236件  
302件  
47件  
15件  
9件

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、そうは思いませんでした。  
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。

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、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。

### テキストデータ

### 数値評価

# テキストマイニングの手順

## 1. データ理解

- 件数や構成比を集計 → データに詳しくなる
  - 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?

## 2. テーマ設定

- 解決すべき課題を選択 → 分析目的を明確にする
  - 明らかにしたい事柄は?
  - 確認したい仮説は?

## 3. テキスト分析

- 課題を解決するため、テキスト分析を実施

# データ理解

- データの概要
  - 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2016年で、下記の10エリアが対象
    - エリアごとに1,000件ずつをランダムに選択

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, 投稿回数

# データ理解

- <https://github.com/haradatm/gssm-201707>

The screenshot shows two GitHub repository pages for the repository `gssm-201707`.

**Top Repository (Branch: master):**

- Commits by `haradatm`:
  - some updated (with a red box around the commit message)
  - ..
- Files:
  - `coding-rule.zip` (with a red box around the file name) - 演習で主として使用
  - `coding-rule_new.zip`
  - `rakuten-all-mac.txt.zip`
  - `rakuten-all-win.txt.zip`
  - `rakuten-all.xlsx.zip` (with a red box around the file name) - 演習で主として使用
  - `rakuten-all_2014-2016.xlsx.zip`

**Bottom Repository (Branch: master):**

- Commits by `haradatm`:
  - first commit
  - ..
- File:
  - `khcoder-slides.zip` (with a dashed blue box around the file name)

**Text at the bottom:**

KH Coder のチュートリアル  
(from SlideShare)

© 2017 GitHub, Inc. [Terms](#) [Privacy](#) [Security](#) [Status](#)

# データ理解

- データファイルの説明

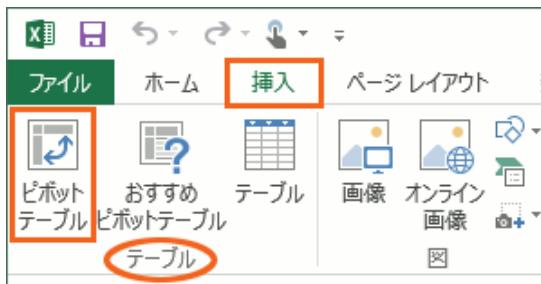
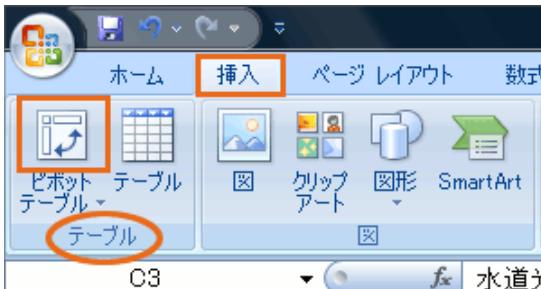
データファイル名	件数	データセット
rakuten-all.xlsx.zip 講義ではこのファイルを 主に使います	10,000	<ul style="list-style-type: none"><li>レジャー+ビジネスの 10エリア</li><li>ランダムサンプリング</li><li>EXCEL 形式</li></ul>
rakuten-all.win.txt.zip ※	10,000	<ul style="list-style-type: none"><li>レジャー+ビジネスの 10エリア</li><li>ランダムサンプリング</li><li>テキスト形式 (シフトJIS, CRLF改行)</li></ul>
rakuten-all.mac.txt.zip ※	10,000	<ul style="list-style-type: none"><li>レジャー+ビジネスの 10エリア</li><li>ランダムサンプリング</li><li>テキスト形式 (UTF-8, LF改行)</li></ul>

※ テキスト形式 (拡張子が .txt のファイル) もあります, Rなど EXCEL 以外で分析する場合に使ってください

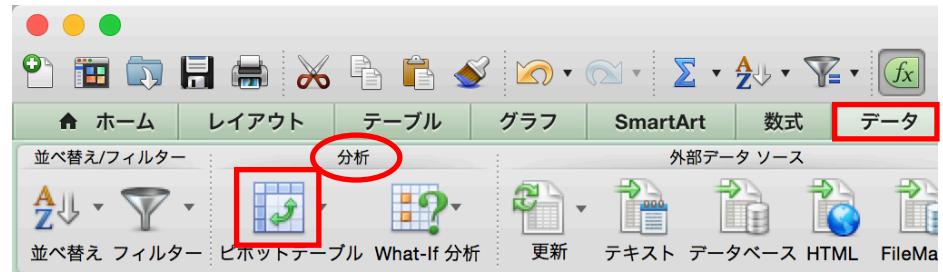
# 演習－データ理解

- ・ピボットテーブル(EXCEL)を使ったデータ集計
  - ・ファイル rakuten-all.xlsx を開く
  - ・A～S列を選択し,ピボットテーブルを作成する

Windows



Mac



【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

【Mac】 Excel 2011

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

# 課題 —データ理解

- EXCEL を使ってデータ集計を行い,発見した特徴でデータセットを説明(要約)する
  - 各人でデータ集計を行う
  - 周囲の4人前後でグループを作る
  - グループ内でデータ集計で発見した特徴を共有
  - グループごとにデータセットの説明を発表

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?

# 論文紹介 —辻井・津田,2012

- 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 / サービ平均	立地	平均 / 部屋	平均 / 設備・設備	平均 / 風呂	平均 / 食事	平均 / 総合
A_レジャー	4.13	4.20	4.02	3.91	4.19	4.20	4.21
B_ビジネス	3.87	4.23	3.92	3.77	3.64	3.85	4.03

- 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



そのため、**数値評価のみから違いを見つけるのは難しい**  
→ レジャーとビジネスでは、**評価すべき項目も異なる**  
→ テキストと対応付ければ、**同じ点数でも差異がある**